Evolution and Engineering in Escherichia coli

Wendel, Sofie

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

Evolution and Engineering in *Escherichia coli*

PhD Thesis

Sofie Wendel

April 2016

The Novo Nordisk Foundation Center for Biosustainability
The Technical University of Denmark

Main supervisor:
Senior Scientist Morten Nørholm

Preface

This thesis is written as a partial fulfilment of the requirements to obtain a PhD degree at the Technical University of Denmark. The work presented in this thesis was carried out between May 2013 and April 2016 at the Novo Nordisk Foundation Center for Biosustainability at the Technical University of Denmark, and included a research stay at AMABiotics at the CIIR institute at Hôpital Pitié-Salpêtrière in Paris, France. The work was supervised by Senior Scientist Morten Nørholm and Dr Susanna Seppälä. Funding was provided by the European union's 7th Framework programme [FP7-People-2012-ITN], under grant agreement no. 317058, Bactory.

The thesis was evaluated by Professor Søren Molin from the Technical University of Denmark, Assistant Professor Johan Rockberg from Stockholm University, and Team Leader Betül Kaçar from Harvard University.

_____

Sofie Wendel

Copenhagen, April 2016

Abstract

Synthetic biology (synbio) is a chance for our societies to advance from oil-dependent to bio-based production, with the help of microbes. The two pillars of synbio are fundamental biological knowledge and applied engineering. In the present thesis, these two aspects of synbio are explored, and their connections described. First, the different definitions and features of synbio are covered, after which two individual research projects are reported.

Evolution is the basis for all life on Earth, and we are constantly learning more about its mechanisms. This thesis describes important elements of evolutionary theory, and in particular the controversy surrounding the theory of adaptive mutagenesis. Furthermore, experimental results showing how adaptive mutation is strongly associated with transcription are reported.

When constructing cell factories, display of proteins on the cell surface by using the cell machinery is an attractive technology. This thesis includes the report of a new bacterial surface display platform that makes use of green fluorescent protein and nanobodies for easy detection and evaluation of surface presentation.

While representing the two pillars of synthetic biology, the two research projects presented also illustrate how every synbio venture contains features of fundamental as well as applied science. With fundamental studies inspiring new engineering efforts, and application development enabling further study of basic biology, it is clear how synbio is created in the overlap between the two.

Dansk resumé

Syntetisk biologi (synbio) er en mulighed for vores samfund at avancere fra oliebaseret til biobaseret produktion, ved hjælp af mikroorganismer. Grundforskning og anvendelsesfokuseret ingeniørvidenskab udgør synbios to søjler. I denne afhandling bliver disse to aspekter af synbio udforsket, og deres forbindelser beskrevet. Først beskrives de forskellige synbio definitioner, og herefter præsenteres to individuelle forskningsprojekter.

Evolution er basis for alt liv på Jorden, og vi lærer hele tiden mere om dens mekanismer. Denne afhandling beskriver vigtige elementer af evolutionsteori, og specielt kontroversen omkring teorien om adaptiv mutation. Afhandlingen indeholder eksperimentelle resultater der viser hvordan adaptive mutationer er stærkt associerede med transkription.

Ved konstruktion af cellefabrikker er overfladeudtrykkelse af proteiner gennem anvendelse af det cellulære maskineri en attraktiv teknologi. Denne afhandling inkluderer en rapport om udviklingen af en ny platform for overfladeudtrykkelse, der bruger grønt fluorescerende protein og *nanobodies* for simpel påvisning og kvantificering af overfladepræsenteret protein.

De to forskningsprojekter der præsenteres i denne afhandling udgør eksempler på synbios to søjler, men viser også hvordan hver synbio venture indeholder aspekter af grundlæggende såvel som anvendt forskning, og at synbio opstår i overlappet mellem de to.

Acknowledgements

I have truly enjoyed the past three years at Cfb. Doing a PhD under these circumstances has indeed been a privilege, and I am thankful for the many exciting experiences I have had, and for being able to refer to them as my job.

Morten, I have never met anyone with your enthusiasm (ah, of course, with one exception…) and I am thankful for you always sharing this with me, as well as always, always being there when needed. You're doing a great job with the open-door policy. Susanna, I have never met anyone as enthusiastic as you about science! Thanks for always spreading joy and knowledge around you. I was very lucky to be introduced into my PhD by you. To all MSB group members – Morten, Ulla, Susanna, Tonja, Karina, Mafalda, Dario, Roberto, Maja, Ida, Emil, Victor, Se Hyeuk, Virginia – a deep thank you for providing a wonderful work environment, and for truly caring for one another. That really means something. Thanks Emil, for great collaborations, and Mafalda, thanks for always being your unique mix of brutally honest and brutally smart.

Dearest Bactories, you have been the best part of this story. I think we made a beautiful group and I am proud of us all. An extra thanks to the Bactory Summer school committee – Klara, Patricia, Henrique – it was a real pleasure being your teammate. Thank you to the many wonderful people at Cfb, and especially to Søren and Mette for making Bactory what it was. I am thankful to Antoine Danchin,

Agnieszka Sekowska and Patrice Garnier for welcoming me to Paris during my industrial stay at Amabiotics.

A very warm and special thanks to the best office ever; I am so glad I didn't move out during those first months when I thought the Southern-European communication style would drive me mad… Seriously – you are all amazing! I have loved being happy, depressed, frustrated and laughing with you.

Finally, I want to thank my family for their support while doing my PhD, and even more for the fact that heir support never had anything to do with me doing a PhD. Last, but not least: Sebastian, thank you for being such a character, and for allowing me to be so as well.

Sofie Wendel
Copenhagen, 2016

List of publications

*Equal contribution

Paper I

Generation of mutation hotspots in ageing bacterial colonies.

Agnieszka Sekowska*, Sofie Wendel*, Morten H.H. Nørholm, Antoine Danchin. (2016)

Manuscript under revision for Scientific Reports.

BioRxiv doi: http://dx.doi.org/10.1101/041525.

Paper II

A Nanobody:GFP bacterial platform that enables functional enzyme display and easy quantification of display capacity.

Sofie Wendel, Emil C. Fischer, Virginia Martínez, Susanna Seppälä, Morten H.H. Nørholm. (2016)

Microbial Cell Factories 15:71, doi: 10.1186/s12934-016-0474-y.

Abbreviations

8-oxoG	8-oxoguanine
AT	Autotransporter
BAM	β-barrel assembly machinery
cAMP	Cyclic AMP
C-deamination	Cytosine-deamination
CRISPR	Clustered Regularly Interspaced Palindromic Repeats
CRP	Cyclic AMP receptor protein
GFP	Green fluorescent protein
GMO	Genetically modified organism
IAC	Inheritance of acquired characteristics
IM	Inner membrane
IMP	Inner membrane protein
LPS	Lipopolysaccharide
mRNA	Messenger RNA
NB	Nanobody
OM	Outer membrane
OMP	Outer membrane protein
OmpA	Outer membrane protein A
RNC	Ribosome-nascent chain complex
scFv	Single-chain variable fragment
SRP	Signal recognition particle
Synbio	Synthetic biology
Tat	Twin-arginine translocation

Table of contents

Introduction and thesis structure

Prokaryotes are according to some calculations the most abundant biomass on Earth [1]. Microbes are able to survive in extremely acidic, hot, wet, and dry environments – in fact, they survive even in space [2, 3]. Additionally, there are at least as many bacterial cells as human in our bodies [4, 5]. What conclusion can we draw from these facts? Firstly, that humans are less independent than we like to think, incapable of surviving without microbes. Secondly, that amazing abilities exist among the microscopic creatures of the world, and that this harbours a huge potential also for us, if we learn from it.

Using the capabilities already developed by the microbial life around us and applying them when faced with new challenges is not new. People have used microbes for production of bread and beer for thousands of years [6]. Using cellular machineries to produce drugs is a major practice in the pharmaceutical industry. It is thus a well-established method for humanity to turn to microbes when faced with a problem in our human world.

One of the largest issues ever faced by mankind, highlighted by several United Nations reports and latest in the Sustainability goals agreed upon by the General Assembly in September 2015, is climate change and the unsustainable way of living that our societies have grown accustomed to [7]. Animal species are disappearing from the face of the Earth at a rate estimated to be up to 100 times higher than it would be without human impact [8]. Average temperatures have risen by 0.8 degrees over the last 130 years and are projected to rise

another 1-3 degrees by year 2100, depending on which efforts are undertaken to halt negative human influence [9]. In a 2015 report, the World Wide Fund for Nature (WWF) describes the terrible state of Earth's oceans, with examples such as record-high acidification due to CO_2 emissions, declines by 49% in marine vertebrate species since 1970, and even more dramatic declines in fish species [10]. World oil consumption rose by 65% between 1973 and 2013, to 3.7 billion tonnes of oil equivalent per year [11]. In Scandinavia, meat continues to be a major food source, and Swedes have increased their meat intake by 40% since 1990, in spite of a reduced meat consumption being one of the United Nation's main recommendations for assisting in reaching the sustainability goals [12, 13].

Faced with these uncomfortable facts, humans are again looking to the microbial world for inspiration. In combination with the advanced genetic engineering techniques developed over the last decades, much effort is put into moving production from fossil fuel-based systems to microbial-based value creation. The modern level of bacterial-based production is part of the new and promising field of synthetic biology.

Thesis structure

In the first chapter of this thesis, I outline the concept of synthetic biology and explain how it is used to increase knowledge of fundamental mechanisms, as well as how it is using acquired knowledge to rebuild existing biological systems.

Next, with *Escherichia coli* as model system, the relevance of evolution for synthetic biology is discussed. In the second chapter, I describe the theory of evolution and the discussion about how randomness, selection and adaptation interact. A detailed understanding of this

aspect of life is a crucial part in the successful synthetic creation of new or improved organisms.

In the third chapter, I outline how synthetic biology can use obtained knowledge of biological systems to develop very specific technologies, in this case particularly microbial surface display. I introduce the biotechnological method of surface display, describe the main microbial machineries, and exemplify how they can be applied for presenting a protein of choice on the cell surface.

The fourth chapter discusses the crossing points between the two main themes of the thesis, and their future perspectives.

Hereafter follows a paper with the findings of a study on how ageing bacteria adapt to their environment. It describes the highly non-random mutational patterns that appear under these circumstances.

The last chapter contains a paper where a new *E. coli* platform for surface display and detection is depicted. This article shows how molecules found in nature can be combined in a synthetic biology approach to develop new useful bioengineering tools.

Chapter 1
Next generation biotechnology

Traditional biotechnology uses methods like cross breeding and selection of individuals with desired traits, and is still used today in food-producing biotech companies since anything associated with GMO (genetically modified organisms) is fiercely rejected by many of the EU member states [14]. Genetic engineering became possible with the development of PCR (polymerase chain reaction) and restriction cloning techniques in the 1970s and 80s, and started a new era in biotechnology, which has seen an impressive advancement of DNA editing, sequencing and synthesis technologies over the last decades. The concept of synthetic biology became a buzzword in the early 2000s, introducing a new way of thinking about biotechnology and molecular biology.

1.1 Synthetic biology and its two pillars

Synthetic biology has over the last decade emerged as one of the really big trends within scientific research, and it has attracted huge amounts in investments for both academic and entrepreneurial activities [15]. In spite of its status as a topic of broad and intense interest, the definition of synthetic biology is somewhat loose.

1.1.1 What is synthetic biology?

The term synthetic biology was in 1974 used by Waclaw Szybalski to describe the "synthetic phase" of biology that he foresaw to be the future [16]. The concept of synthetic biology (often abbreviated

synbio) has evolved since then and is today defined somewhat differently by different people. Some say it is exclusively the building of formerly non-existing life forms, and define the steps prior to that, involving understanding the natural biology better, as systems biology [17]. Others, however, bring together the understanding of how a living system is built and the eventual re-building of it in their definition of synthetic biology [18]. This view is inspired by physicist Richard Feynman's famous quote: "What I cannot create, I do not understand." According to Drew Endy, who was instrumental in establishing the term's usage when he used it for the International Genetically Engineered Machine (iGem) competition, synthetic biology is a method, an approach to making new applications that builds on genetic engineering but adds automated DNA synthesis, standards and abstraction to the process [19]. While some are clearly separating the engineering field of synbio from the basic understanding that is systems biology research, others emphasize the understanding of how a system works as a fundamental part of synthetic biology.

1.1.2 Synbio accomplishments and perspectives

The last decade has seen an explosion in synthetic biology-based inventions and technologies, resulting in a whole new community of academic researchers, entrepreneurial start-ups, and do-it-yourself biologists. Start-ups have emerged in large numbers from particularly California and the UK, with new entrepreneurs aiming at applying synthetic biology for a great variety of purposes. Ginkgo Bioworks makes designer roses, while Caribou biosciences does gene editing and Bento bioworks creates 3 kg-heavy do-it-yourself laboratories. Amyris and Sanofi have received much attention for their yeast-produced anti-malarial drug Artemisinin, and Oxitec is the object of much attention due to their genetically modified insects,

prospected to be released in nature to eradicate disease-carrying mosquito species [20]. In Denmark, microbe-based production systems have long had a steady base with insulin-producing Novo Nordisk, enzyme-making Novozymes, and food-focused Chr. Hansen. Academic research has set the stage for the explosion of entrepreneurial successes, and includes genetic tools development, metabolic engineering advances, and natural products manufacturing [21]. DNA synthesis and sequencing have seen increased speed and reduced costs, and genome editing has reached entirely new levels with the development of MAGE (Multiplex Automated Genomic Engineering) and CRISPR (Clustered Regularly Interspaced Palindromic Repeats) techniques. Additionally, synbio is testing new ground when it comes to knowledge sharing. In an industry traditionally based on exclusive rights and patenting, parts of the synbio community are advocating a new *open biology* and sharing strategy that challenges old structures, much in resemblance with the open source development within computer science [22].

1.1.3 Synthetic biology from two perspectives: evolution and engineering

Within the concept of synthetic biology exists an interesting conflict of principles (Figure 1). From a biological perspective, the process of evolution as a means for new functions and forms to emerge constitutes the basis of life. From an engineering perspective, however, this sort of behaviour of a system is inherently problematic and such appearances should rather be controlled and designed [23].

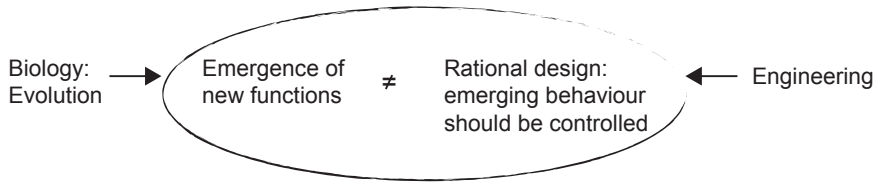


Figure 1. The synbio conflict. Biology and engineering have principally different views on evolution.

At the same time, synthetic biology has been stated to be "extremely primitive due partly to fundamental inadequacies in our understanding of biological circuit design" (Michael B. Elowitz in [17]), and design of evolutionarily robust biological circuits was pointed out as a synbio safe point in a priority report presented in 2009 [24]. As an extreme standpoint, for biological systems to fully fulfil engineering requirements, the only evolution taking place should be programmed according to Smolke and Silver [25]. This conflict highlights the existence and importance of two sides of synthetic biology; the need for understanding fundamental mechanisms of biology in order to engineer life, and the increased understanding of fundamental mechanisms that comes from engineering the same system. The two are intimately linked, with tools developed for synbio being applied for fundamental research, and fundamental research continuously feeding in to synbio. Exemplified by the case of evolution, this could mean exploiting evolutionary processes in synbio constructions.

What can be extracted from the existing variety of synbio definitions is that synthetic biology requires understanding of life to a level where we can rationally design and build it *de novo*. Whether thinking about synthetic biology as a method or a field, iteration over

cycles of building, and studying and learning from the results thereof, is essential.

In this thesis, two complementary paths of synthetic biology are explored: evolution is studied to understand mechanisms that can then be applied for rational design, and engineering using known mechanisms is performed, creating tools that help deepen understanding of said mechanisms as well as improving cell factories.

1.2 Cell factories

The concept of cell factories is central to most synthetic biology applications and technologies (the famous exception being an *in vitro* system incorporating full gene circuits on a piece of paper [26]). As the name implies, a cell factory is using the machinery of a cell to produce a compound of interest, e.g. a protein, a medicinal molecule, or a food ingredient. While yeast cell factories have been used for centuries for production of alcohol and bread, the development started taking big leaps forward in the 1980s, with the introduction of advanced genetic engineering techniques.

1.2.1 Principle and principal benefits of cell factories

Today, oil is the raw material behind almost everything we use in our everyday lives: the paint on our walls, the plastic in the containers we use, the fabric of our clothes, and even most of the vanilla taste in our cakes [27]. One of the major benefits of cell-based production is the opportunity to make molecules we need without petroleum, thereby reducing carbon dioxide emissions and environmental damage. Cell factories, in contrast to common industries, use sugar in various forms as raw material. Through metabolic rewiring or boosting of enzymatic pathways, cells can be tuned into producing a great

variety of products, and in the best cases use renewable resources in efficient ways (Figure 2).

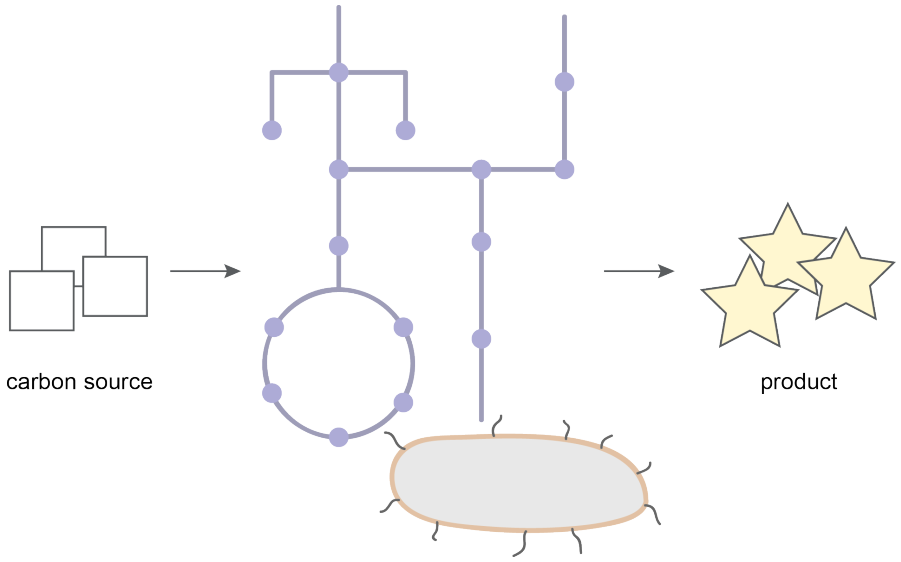


Figure 2. The principle of a cell factory. Cheap, naturally occurring sugars (carbon source) are fed to microorganisms, whose metabolic networks are wired to allow production of valuable compounds such as drugs, nutritional molecules, or fuels (product).

1.2.2 Organisms and products

Many different organisms have been explored as cell factories. The main workhorses in synthetic biology are *Escherichia coli* and *Saccharomyces cerevisiae*, but also mammalian cells for particularly pharmaceutical proteins [28]. Many alternative organisms are continuously being explored, and moss, algae, insect cells and anaerobic bacteria are a few of the more exotic ones. Eukaryotic organisms are preferably used for production of proteins requiring post-translational modifications [29], but oftentimes bacteria are used due to their simplicity of modification, low cost and fast growth [30]. While it is proven time and time again that it is possible to make all

sorts of complex molecules in microbes, in many cases making it at high enough titers at low enough prices remains an issue. Fully realising the potential of synthetic biology and cell factories will require collaboration between disciplines, academia and industry, as well as deepened understanding of cellular mechanisms [17, 31].

1.3 The model organism

E. coli is often referred to as the most well characterised species on Earth. It has been used by humans for a century and provided clues to many of our questions about life and our own bodies' molecular details [32]. Some strains of *E. coli* are pathogenic, whereas others exist in the normal human and animal gut flora. Its benefits as a host for synthetic biology-based engineering and production include a fast doubling time, with cells dividing every 20 minutes under optimal conditions, easiness to grow to high densities, and a vast array of available methodologies for its genetic manipulation [33, 34].

While *E. coli* may in everyday life be mostly associated with gut infections, *E. coli* is really a saviour rather than a foe. As one of the most important model organisms there are, *E. coli* has taught us the basics of bacterial conjugation [35], gene regulation [36] and gene structure [37]. Research in *E. coli* has laid the foundation for much of modern biotechnology, with it being essential in the development of recombinant DNA technologies [38]. The use of *E. coli* for recombinant protein production is widespread, and that for purposes such as pharmaceuticals, biofuels and amino acids manufacturing [39]. A famous example is the production of insulin that was accomplished in the 1980s [40]. Today, *E. coli* continues to be one of the most used organisms within biotechnology and synthetic biology. Drawbacks of using *E. coli* include the afore-mentioned incapacity to

produce post-translational modifications, and production of endotoxins [30].

1.3.1 Different strains

Although *E. coli* is one species, there are many, very different, strains. The two major laboratory strains are the so-called B and K strains. Both strains have the typical circular *E. coli* chromosome encompassing 4.6 million basepairs coding for around 4500 proteins.

The work in this thesis was performed using the K strain K12 MG1655 and the B strain BL21, for the study of evolution and mutation mechanisms as described in Chapter 2 and Paper I, and for the purpose of surface protein production as described in Chapter 3 and Paper II, respectively.

K12 MG1655 was the first *E. coli* strain, and one of the first genomes ever, to be fully sequenced [41], whereas BL21 was sequenced in 2009 [42]. Comparison of their genomes revealed differences mainly in their membrane constitution: BL21 cannot form capsule polysaccharides and also lacks fully functional lipopolysaccharide (LPS), due to a truncation in the core oligosaccharide [42]. Both strains are routinely used in laboratory work, with BL21 being especially suitable for protein production, partly due to its lack of certain proteases [34].

Chapter 2
The adaptation of life to living

Evolution – this beautifully simple yet refined basis for life the way we know it. While the basic principle seems so easy to grasp, there are layers of evolution that are far more intricate. Do subtypes of evolutionary mechanisms exist and can they be turned on and off? Can cells sense environmental cues and use them to steer their own evolution? These questions might sound bold, but one could argue that it would be even bolder if life did not develop high degrees of control of the most important process of all, and recent research indeed reveals that such mechanisms exist.

2.1 The basic principles of evolution

We may think that we all know and agree on what evolution is and how it works, but when it comes to discussing the details of it, it turns out there are many controversial discussion points. The 130-year-old concept of evolution as described by Darwin has been challenged by many differing ideas with regards to the ever-fascinating question of heredity and development of life.

2.1.1 Instructed evolution

In 1809, the French biologist Jean-Baptiste Lamarck was the first to present an idea of evolution of life; he suggested that life on Earth did not appear there exactly the way it presently looked, but that it had developed over time [43]. Lamarck's idea, not at all controversial at the time, was that organs that were much used by an animal would

grow bigger, whereas organs that were not used would become smaller. These advantageous changes would take place in individual animals during their lifetime, and be transferred to their offspring. Life, he argued, is always developing towards increasing complexity. Although Lamarck was a pioneer who suggested that life was not fixed but changed depending on its environment, something that was heavily criticised by religious representatives, he is remembered for the later ridiculed theory that the environment instructs evolutionary changes. Interestingly, the Lamarckian evolution-thought that acquired characteristics can be transferred to the next generation is today gaining new relevance. In CRISPR research as well as in the fields of epigenetics and that of evolution of life, where IAC (Inheritance of Acquired Characteristics) is the term used to describe Lamarckian-style evolution events, the topic is attracting renewed interest. A discussion of some of these cases is placed in the last section of this chapter.

2.1.2 Natural selection

In the mid 1800s, Charles Darwin's model of evolution, which still holds today, was presented [44]. Building on among others the work of Lamarck, both Darwin and Alfred Russel Wallace, independently, developed theories of evolution based on what Darwin called *natural selection*. Inspired by the field of economics, where the pressure of population growth on a state's capacity to withhold welfare for all had been pointed out, Wallace and Darwin concluded that the same must be true for the animal kingdom. Due to limited food or space, a population could not grow freely and continue to expand indefinitely. If an individual had traits that were beneficial under the circumstances it was living in, it would have an increased likelihood of surviving and propagating. With each generation, the beneficial trait would become more widespread in the population, and in the

long run entirely new functions could develop [44]. Natural selection was thus defined as the principle that individuals with beneficial traits, or in other words with the highest *fitness*, would have the best chance of being preserved, and that they would pass on their characteristics to their offspring.

At first, natural selection was argued for with observational data based on bird species studies, and the Lamarckian theory which had its base in "folk wisdom" did not disappear easily [45]. With the discoveries of genes and mutation in the 19th century, however, Darwin's theory became the dogma. When the details of genetic inheritance were revealed in the early 1900s, this new knowledge was combined with Darwin's ideas of evolution through natural selection into what is commonly referred to as neo-Darwinism or *modern synthesis*: evolution of a species due to natural selection of advantageous, randomly generated mutations from a genetically diverse population. The Luria-Delbrück fluctuation experiment, reported on in 1943, was instrumental in its demonstration of how mutations in *E. coli* occurred before exposure to a specific environmental pressure, and hence were not due to instructed adaptation [46]. A third component of evolutionary theory is the concept of genetic drift that was first described by Sewall Wright; randomly generated mutations' random persistence in a population (Figure 3) [45, 47].

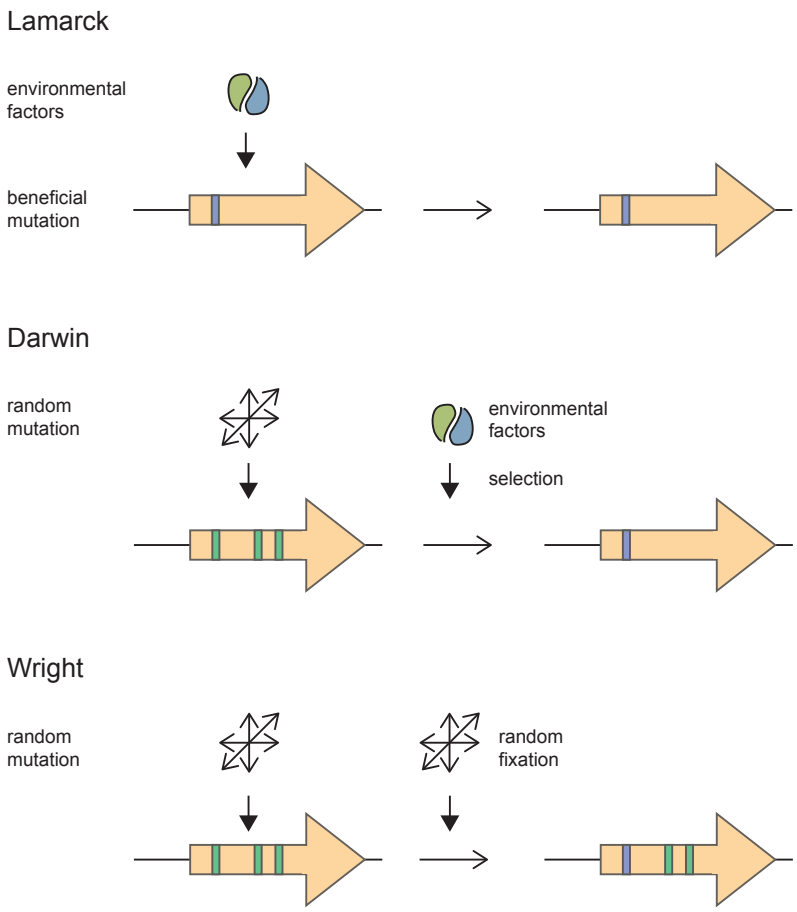


Figure 3. Schematic overview of evolutionary concepts. In Lamarckian theory, environmental factors direct specific, beneficial mutations that allow an organism to adapt. The Darwinian principle is that of randomly generated mutations enabling selection of beneficial mutations when an organism is subjected to a selection pressure caused by environmental factors. The Wrightean module of evolution concerns randomly generated mutations that are fixed in an organism through random processes. Adapted from [48].

2.1.3 Terms and definitions related to evolution mechanisms

Shortly after consensus was reached about the fact that life evolves

and that it does so via natural selection, discussion was sparked anew on the mechanistic details of evolution, and the model organism *E. coli* once again became the focus of attention. A particularly debated topic is that of adaptive evolution, i.e. the way mutations appear in microorganisms adapting to a new environment. The discussions taking place revolve around numerous critical terms presented in Table 1.

Table 1. Overview of commonly used terms in the discussion of evolution and adaptation. Partly adapted from [49] and [50].

Term	Definition
Directed mutagenesis	Under selection, only mutations that relieve the selection pressure appear (no unselected mutations).
Adaptive mutagenesis	Under selection, mutations that relieve the selection pressure appear, with or without additional, unselected mutations.
Stress-induced mutagenesis	A general mutagenic state induced in response to stress.
Transient mutagenesis	A state of temporarily increased mutation rate affecting the whole genome.
SOS response	DNA-damage stress response that upregulates ca. 40 genes involved in DNA repair and DNA damage tolerance.
RpoS response	Large general stress response driven by the transcriptional activator RpoS, upregulating transcription of ca. 340 genes.
Hypermutagenesis	A state of extremely high mutation rate accumulating mutations across the whole genome.
Stationary phase	The stage of a laboratory batch culture when one or more nutrients are finished. Accumulation of waste due to cell metabolism.

Quiescent state	Non-dividing, dormant state of a cell where it does not replicate.
Starvation	Lack of one or more nutrients required for growth.
Adaptive evolution	Evolution through selection for improved fitness.

2.2 The controversy of adaptive mutation

It is remarkable how much controversy the studies of cells' adaptation to their environment have ignited in the scientific community. Two distinct camps exist, and have been fiercely counter-arguing each other for many years. One side determinedly states that all observations on adaptation can be explained by classic, neo-Darwinian selection of pre-formed mutations and growth advantage. Other researchers interpret the same scientific results as evidence for stress-induced mutagenesis restricted in time and genomic space via a multitude of molecular mechanisms. The heat of the debate can be sensed via expressions such as "we *feel* that there can be no real controversy" appearing in scientific articles on the topic [50]. What, then, is the controversy about?

2.2.1 The lac system

In two papers published in 1988 and 1991, John Cairns and colleagues, particularly Patricia Foster, reported experiments that have since formed the basis for a large number of studies meant to decipher how bacterial cells adapt to their environment [51, 52]. *E. coli* cells with a chromosomal deletion of the *lac* operon, containing the genes for lactose fermentation, were plated on lactose-containing medium. The cells harboured an F' conjugative plasmid with the *lacI-lacZ* genes disrupted by a leaky +1 frameshift mutation, causing reduction of LacZ activity to 2% of that of revertants [53]. While the

cells did not form any colonies immediately after being plated on lactose, colonies started to appear after two days of incubation and continued to form over the following eight days. The papers concluded that the bacteria were sensing and adapting to their environment, and that "cells may have mechanisms for choosing which mutations will occur" [51].

After the original experiments by Cairns and Foster, the *lac* system was extensively explored as many researchers repeated the experiment or set up similar tests, but without the scientific community ever reaching a conclusion satisfactory to everyone. Designing a robust system for studying these types of mechanisms is evidently very complicated. In this long-lasting debate, with the latest review coming out last year [54], the controversy has mostly been about the interpretation of the scientific results, rather than about the results as such. Adaptation, defined as "the process by which new mutations arise under selective conditions regardless of the mechanism" [55, 56], *is* taking place, the question is just *how*? A number of different models for explaining the observations have been put forward and are presented next. The first two propose stress-induced mutagenesis to account for the adaptive mutations, whereas the latter two refer to growth and selection.

The directed evolution model

One model states that mutation is directed; under stress, there is an increase in overall mutation rate in all cells and based on sensed cues from the stressing environment, mutations are specifically directed to genes that would be beneficial to mutate. This was the original line of thinking in the very first *lac*-system studies but has not received much support [51, 52, 57].

The hyper-mutable subpopulation model

A second model suggests that a small subpopulation of the plated cells become hypermutators in response to stress [58]. Beneficial traits are then selected from this subpopulation. Evidence in support of this model include a generally higher mutation frequency in *lac*+ cells [59–61] whereas the counterarguments are that the mutation rate required to obtain that many mutations would be unrealistically elevated [62].

The amplification model

This model refutes stress-induced mutagenesis and explains the Cairns-Foster experimental results by local amplification of the leaky *lac* allele [54, 56]. Continuous amplification in this model leads to firstly a growth advantage, and secondly an increased probability of mutagenesis simply due to the sheer number of *lac* alleles being higher. Foster, on the other hand, argues that if the amplification model were true, the increase in mutants should be exponential rather than show the observed linearity [49].

A compromise model?

As late as last year, Maisnier-Patin and Roth discussed the controversy again, and proposed a new model that combines characteristics of the previously suggested ones [54]. They propose that prior to selection, the F' plasmid copy number has increased in some cells, via unknown mechanisms. In the plated, non-growing cells, energy from the leaky *lac* allele is then used to replicate the plasmid, and each such event is a chance for the mutant *lac* gene to revert. At the same time, both the SOS response and the RpoS response are activated. The error-prone DNA polymerase DinB encoded on the plasmid is responsible for increasing reversion rate for *lac*, while not affecting the chromosomal DNA since the

chromosome is not replicating. They conclude that the mutations are directed in the sense that only the plasmid is increasingly replicated in the presence of increased levels of DinB, and that once a reversion takes place, all non-revertants are lost due to selection, leading to direction of mutagenesis to those nucleotides that limit growth. In essence, they are arguing how selection mimics any sort of induced mutagenesis, and that the dogma that mutation and selection are independent still holds true.

While some researchers have been discussing this topic for decades, it is noteworthy that others strongly criticise the whole experimental set-up and go as far as to say that the whole Cairns-Foster system debate is a result of "serious hubris" and a waste of time as well as a mistake to ever be published [63]. It is indeed not very bold to say that the *lac* system is somewhat flawed, and that a different experimental set-up with more easily interpreted output would be beneficial for the discussion.

2.2.2 Transcription-associated mechanistic models

In the late 1980s and mid 1990s, two articles reported non-Lamarckian, transcription-related hypotheses for adaptive mutation [64, 65]. During transcription, in the transcriptional bubble, the DNA strands are unusually exposed, increasing the risk for DNA damage. This is true particularly for the non-transcribed strand, to which no messenger RNA (mRNA) molecule is base-pairing. Based on this fact, Davis suggested that during transcription in non-growing cells, the non-transcribed strand is prone to mutagenesis and that this could be the mechanism behind adaptive mutation [64]. Bridges reported a related but contrasting model, which has been termed retromutagenesis [65].

Retromutagenesis is the principle of mutation to the transcribed strand during transcription, leading to a beneficial phenotype change that allows growth and thereafter DNA replication and fixation of the mutation also on the non-transcribed (coding) strand and in the genome (Figure 4). Retromutagenesis hence allows for mutation to genes that are transcribed, and in that sense has parallels to the *lac* amplification model, where amplification of a gene allows for its increased mutation. The unique characteristic of retromutagenesis is the reversed order that a mutation occurs in; that a modification is fixed in the DNA first after expression of the mutated gene. Recently, new evidence was presented in support of the retromutagenesis model: mutagenized DNA was fixed by compensating mutation highly dominantly taking place on the transcribed strand [55].

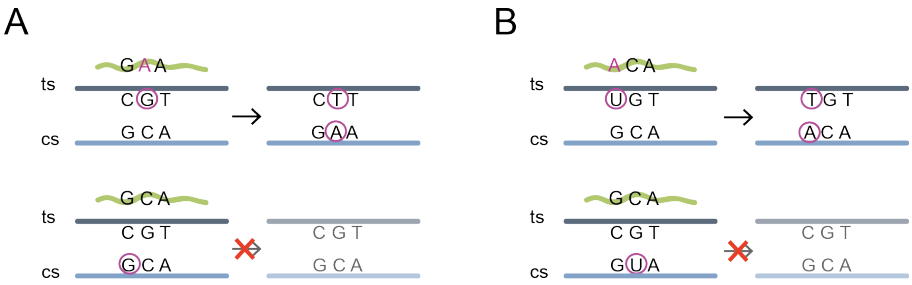


Figure 4. The retromutagenesis principle exemplified with two common mutational mechanisms. (A) Upper panel: Illustration of the mechanism of retromutagenesis after 8-oxoguanine (encircled G) formation on the transcribed strand (ts, dark grey line) that enables insertion of an adenine ribonucleotide in the mRNA (green waved line). The mutated mRNA enables growth and a round of replication that permanently fixes the mutation on the coding strand (cs, light blue line) in daughter cells. Lower panel: 8-oxoguanine formation on the coding strand does not transfer to daughter cells. **(B)** Illustration of retromutagenesis after cytosine nucleotide deamination into uracil (encircled U), enabling base pairing with A. Upper

and lower panels show result of cytosine deamination on transcribed and coding strand respectively, as detailed in (A). Figure from Paper I [66].

The fact that the adaptive mutation controversy has persisted for almost 30 years shows with no room for doubt that the Cairns system is imperfect and that other ways of studying adaptation would be helpful. One lead in the hunt to understand how organisms adapt to their environment might be provided by the retromutagenesis model, but the systems where it has been tested until now are dependent on mutagens. To complement the existing results, we developed a new experimental system for studying adaptation of ageing *E. coli* cells to their environment.

2.3 The adenylate cyclase – CRP experimental system

In an attempt to shed more light on the process of adaptive evolution, part of the work for this thesis was done on a completely new experimental system, designed to address some of the questions regarding how bacterial cells adapt to their environment. The desire for such a system is for it to be minimally perturbed compared to the wildtype, in order to have any general implications, while at the same time generate enough mutants to enable their study; hence, for it to have a weak mutator phenotype.

2.3.1 Adenylate cyclase and the cAMP receptor protein

One of the main regulators in *E. coli* is the *cyaA*-encoded kinase adenylate cyclase. This enzyme catalyses the conversion of ATP into cyclic AMP (cAMP), one of the main second messenger molecules in *E. coli* [67]. Next, cAMP binds to the cAMP-receptor protein (CRP, also known as Catabolite activator protein, CAP) and causes it to change conformation, which activates it as a transcriptional regulator controlling expression of more than 100 genes (Figure 5) [68].

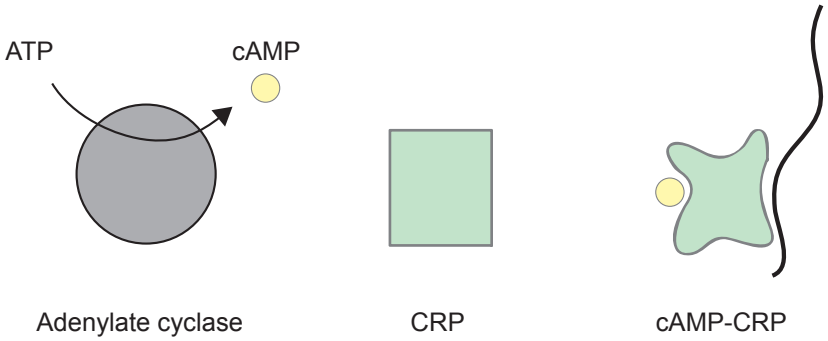


Figure 5. Principle of cAMP-CRP transcriptional regulation. Adenylate cyclase converts ATP into cAMP, which upon binding to CRP causes a conformational change in CRP, leading to its activation as a transcriptional regulator.

2.3.2 Observation of cells' adaptation over two months

Paper I of this thesis describes a study where we sequenced almost 100 genomes of cells that adapted to an inaccessible nutrient over the course of two months. The experimental system is based on the *E. coli* strain K12 MG1655. The *cyaA* gene is removed, which hinders formation of cAMP and consequently formation of the cAMP-CRP complex. Thus, transcriptional regulation of a multitude of genes is inhibited. Furthermore, the strain has a deletion of *fnr*, which codes for a CRP homolog [69].

When the strain is plated, from early stationary phase cultures, onto plates containing lactose-free MacConkey medium supplemented with maltose as the only available carbon source, they grow into small, white colonies but not beyond. However, if the plates are left in the incubator for some days, red protrusions, called papillae, start to form. These are cells that are adapted to their new environment and capable of metabolising maltose. This situation would in principle be similar to a starvation situation in nature, and the

question be: which are the mechanisms behind cells' adaptation to a new, hostile environment? These adaptive mutants continue to form over the course of two months. The goal of the experiment reported in Paper I was to study these mutants in detail to a level that has not been done before, by sequencing the whole genome of a large number of the papillae. Another unique factor of this experiment was the long time period over which the adapted mutants were collected. Each papilla that appeared was purified and characterised phenotypically, and a representative group consisting of 96 of the mutants was sequenced. In addition, Paper I contains an analysis of a larger group of CRP sequences from mutants. A summarising overview of the observations found in the genome sequencing data is found in Table 2.

Table 2. Overview of mutation events in the adaptive mutation experiment of Paper I.

Number of genomes sequenced	96
Number of CRP genes sequenced	594
Number of mutations per genome, average	5
Number of missense mutations per genome, average	2.5
Range of number of mutations per genome (missense)	1-14 (1-7)
Total number of genes with missense mutations	83
Fraction of all missense mutations consistent with 8-oxoG mechanism (CRP)	69% (28%)
Fraction of all missense mutations consistent with C-deamination mechanism (CRP)	28% (66%)
Fraction of 8-oxoG mutations on transcribed strand (CRP)	84% (98%)
Fraction of C-deamination mutations on transcribed strand (CRP)	93% (99%)

The results show a highly non-random pattern of mutation, with adaptive mutation taking place in specific hotspot genes, in a

background of low mutation rate, appearing over the whole time period and showing a strong bias for the transcribed strand. As could be expected, the major hotspot gene is *crp*. Another oftentimes-mutated gene is that of the transcriptional activator (sigma factor) protein RpoS, controller of a general stress response.

The strand bias is one of the most striking results we obtain and the basis for determining on which strand a mutation takes place is the following: Under starvation conditions there are two types of mutations that are widely dominating over all others, and they are functioning through 8-oxoguanine formation and cytosine deamination, respectively [70–72]. In Paper I, all mutations consistent with the typical 8-oxoG or C-deamination mechanisms are assumed to be caused by these mechanisms, and the strand is determined by backtracking the mutation: a C-to-A transversion on the coding strand is consistent with 8-oxoguanine formation on the transcribed strand, whereas a G-to-A transition on the coding strand corresponds to a cytosine deamination event on the transcribed strand (Figure 6a-b). By analysing the genome data with these mutational glasses on, we discovered that, unsurprisingly, 97% of all missense mutations were formed by these common mutagenesis mechanisms, and – more surprisingly – that 84 and 93% of 8-oxoG and C-deamination events respectively were formed on the transcribed strand. This strongly indicates that adaptive mutations are formed in a transcription-associated manner and supports the retromutagenesis model.

8-oxoG mutagenesis leads to transversion mutations (changes between purine and pyrimidine bases), in contrast to the transitions (purine to purine or pyrimidine to pyrimidine changes) caused by C-deamination (Figure 6c). Transition mutations are twice as likely to happen as transversions, and can be caused by tautomerisation

during replication [73]. Distinctive from this, we observe transversion mutations to be dominating, and in addition this domination is increasing with time over the 2-month experimental course. This suggests that the oxidative environment becomes increasingly influential over time. Interestingly, the increase in 8-oxoG mutations is paralleled by an increase in RpoS mutations, something that may suggest that RpoS counteracts 8-oxoG formation. In consistence with this hypothesis, starved *Pseudomonas Putida* cells deficient in RpoS show significantly increased mutation rate [74].

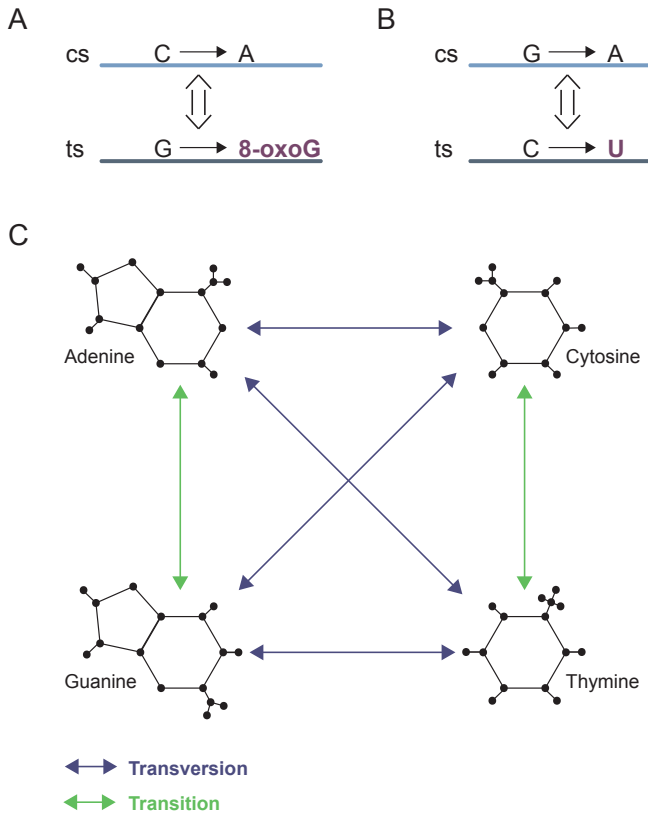


Figure 6. Illustration of mutation principles. (A) Observation of a C-to-A transversion on the coding strand (cs) corresponds to a G being oxidised on the transcribed strand (ts). **(B)** Similarly, a G-to-A transition on the coding

strand corresponds to deamination of C to U on the transcribed strand. **(C)** Transitions are changes between one-ring pyrimidines or two-ring purines, whereas transversions are conversions from pyrimidine to purine or vice versa.

Interestingly, several of the hotspot genes identified in our study are subject to autoregulation, meaning that when their levels decrease, their transcription is triggered [75, 76]. In a 2015 study, Franchini *et al* reported unaffected transcriptional levels of *crp* in a *cyaA* deletion strain grown under glucose limited conditions [77]. 192 genes were differentially expressed in the *cyaA* deletion strain, 152 of those down-regulated as an effect of the absence of cAMP-CRP. Hypothetically, the lack of effect on *crp* expression levels, keeping the gene transcribed under difficult conditions, can be beneficial.

In our experimental set-up, the process is undoubtedly that of selection, but not of randomly generated mutants. Rather, we suggest that adaptive mutations appear in regions that keep being transcribed as cells age (for instance, autoregulated genes), via retromutagenesis and selection of beneficial phenotypes, which then fix mutations in the genome through replication. Autoregulation of genes is a potential mechanism to ensure that adaptive mutation takes place only in specific genes when cells starve and age.

2.4 A new view on evolution and its role in synthetic biology

In the heat of discussing evolutionary principles in general, and adaptive evolution in particular, might it be time to update our views on evolution once again? Perhaps, as is often the case, the true picture lies somewhere in-between the extremes.

2.4.1 A continuum of evolutionary mechanisms

The era of genomic analysis has enabled insight into the molecular mechanisms of mutation and heredity that casts new light on evolutionary processes. Interestingly, the hottest scientific topic of the last years, the CRISPR-Cas system of bacteria, has again brought attention to Lamarckian mechanisms of evolution.

CRISPR (short for Clustered Regularly Interspaced Palindromic Repeats) together with the Cas (CRISPR-associated) nucleases carry out the adaptive immune response of bacteria. By cutting out pieces of invading viruses' DNA and saving them in CRISPR arrays in the bacterial genome, the bacterium can on a later occasion recognise the invader and guide a Cas nuclease to its DNA, where Cas creates a double-bond break [78]. There is a clear Lamarckian/IAC tone to this environment-instructed genetic change that is passed on to the next generation. It is most clearly revealed in the *E. coli* type I-E CRISPR-Cas system, which recently has been shown to have a strong bias for foreign DNA over own DNA, thus fulfilling the important Lamarckian requirement of only environmentally induced genetic changes that are *advantageous* to the organism being inherited to the next generation [79, 80].

The results of Paper I of this thesis suggest a non-random generation of adaptive mutations. There are Lamarckian notes to the observations, albeit not as strong as for CRISPR. Koonin and Wolf (2016) suggest that, rather than limiting our explanatory models to Darwinism *or* AIC, evolution should be seen as a continuum of processes, encompassing Darwinian, Lamarckian and genetic drift modules [80]. From this perspective, one can argue that having a completely random approach to mutagenesis under stress and starvation would be wasteful for the cell. From an energy-efficiency

point of view, it makes sense for there to be a hierarchy of genes that are mutated under stress, starting with transcriptional regulators, rather than for a cell to sample all possible, randomly generated mutations. Evolution, it seems, is not always the process of neo-Darwinian random mutagenesis and selection, but uses a complex web of mechanisms containing also the modules of AIC and genetic drift.

2.4.2 Evolution and synbio

As mentioned in the introduction, the desire for control that is fundamental in engineering of biological systems conflicts with the inherent capacity of living systems to change and evolve. Perhaps, though, it is possible to take the opposite standpoint, and harness organisms' adaptation capacity in synbio constructions. As pointed out by Christina D. Smolke in a discussion paper, "mutation and evolution were widely viewed as an obstacle […], whereas newer approaches are beginning to exploit this unique aspect of biological systems and to design for evolution and adaptation" [17].

For genes to evolve a capacity to evolve is a fundamental paradox, but may still be true. There are in fact several systems, such as error-prone repair, that through history have evolved in spite of their functions as mutation-causers [50, 81]. Our Paper I study points to certain mutational mechanisms being dominant at different time points, and the mutagenesis mechanism is strongly linked to the precise DNA sequence. This highlights the potential to use knowledge of evolution mechanisms for design of genes and organisms to either resist or exploit adaptive mutation. When we understand *how* DNA mutates, we can use it: e.g. in a process where 8-oxoG can be assumed to be a dominating mutation mechanism, varying the G content of individual codons can be a method to

increase or decrease *evolvability* of a gene. One might speculate that nature has done this with important genes, or that 8-oxoG is a mechanism developed by nature to increase diversity over time. Further, although seemingly very different, knowledge of adaptive changes in ageing, stressed bacterial cells might be relevant for eukaryotic ageing and tumour biology [50].

Chapter 3
From observation to creation – putting biological knowledge into practice

The ultimate goal of synthetic biology is to build *de novo* biological systems that are capable of carrying out whichever tasks we have designed them for. Until we reach that point, much of synthetic biology revolves around modifying and adjusting existing biological entities to serve our needs. One such application is making use of the cell machinery for the construction of whole-cell catalysts using surface display technology.

3.1 The different compartments of an E. coli cell and why to go there

While *E. coli* was historically seen as a unicellular prokaryote lacking any compartmentalised structures, this view has changed considerably. The *E. coli* cell has multiple marked off compartments: the inner and outer membranes, the periplasm and the cytoplasm. Additionally, there is a high degree of organisation of protein networks and complexes that carry out various tasks, for example proteins involved in cell division and cell shape maintenance, or machineries for protein folding and secretion [82]. Taken together, the bacterial cell is far from the once used description *a bag of enzymes*.

3.1.1 The cell and its membranes

E. coli is a 2 μm long, rod-shaped bacterium (Figure 7). Being a Gram negative, it has two membranes, inner (IM) and outer (OM), separated by the periplasmic space and a peptidoglycan layer. Together they protect the bacterium from the outer milieu, and give the cell rigidity. The inner and outer membranes differ in their composition, with the two IM leaflets having the same composition of mainly phosphatidylethanolamine (PE) lipids [83, 84]. In contrast, the OM's leaflets differ from each other. The inner leaflet looks like the IM leaflets, whereas the outer leaflet contains endotoxins also known as lipopolysaccharides (LPS). These consist of lipid A molecules linked to oligosaccharides and variable O-antigen polysaccharides, and function both as stabilisers of the membrane and immune response activators in many animals [85].

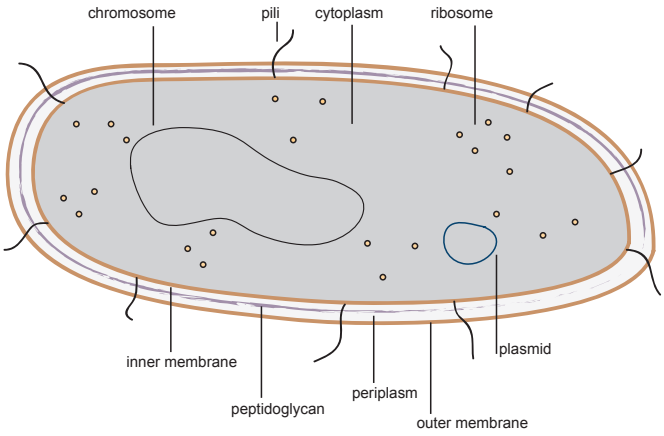


Figure 7. The *E. coli* cell. Important compartments are marked, and include the inner and outer membranes separated by the periplasm and its peptidoglycan layer, chromosomal and plasmid DNA contained in the cytoplasm together with ribosomes, and pili protruding from the cell surface.

In addition to lipids and LPS, *E. coli*'s membranes consist of large numbers of proteins. Membrane proteins constitute an immensely interesting group of proteins, responsible as they are for crucial functions like transport, adhesion and virulence. In fact, genomic data analysis predicts that around 25% of the genes in any organism code for integral membrane proteins [86]. Their importance is illustrated by the fact that more than 50% of all drugs target membrane proteins [87, 88]. In the minimal genome recently presented by the Venter lab, 18% of the 473 genes were membrane-related [89]. One synbio effort that makes use of membrane proteins is that of surface display.

3.1.2 Surface display

In surface display technology, specific membrane proteins are employed for placing a passenger, typically a protein with enzymatic activity, on the outside of the cell, anchored to the cell membrane. The principal benefits of this technology include simplified down-stream processing due to easy access to enzyme and product, potential circumvention of toxicity from intermediates or product, and a highly useful link between phenotypic and genotypic characteristics. Phage display, fusing a peptide to the phage's coat protein and thereby obtaining its co-display on the surface of the phage, became a standard technique for display of recombinant polypeptide libraries after it was first reported in 1985 [90]. Applications are centred on the study of protein-protein interactions with screening of antibody libraries for the identification of highly specific therapeutic antibodies, a key technique in the pharmaceutical industry [91]. Bacterial systems, using both Gram-positive and Gram-negative species, have subsequently been developed and used for purposes such as vaccine development, antibody epitope mapping and bioremediation [92, 93]. A major focus of bacterial surface

display research is to construct whole-cell catalysts by displaying enzymes that can act on various substrates interesting for industrial applications [94].

A recurring difficulty with surface display is the detection of surface-presented protein, and this problem is what we addressed in Paper II of this thesis. We fused the displayed protein to a type of antibody molecule specific for green fluorescent protein (GFP), which allowed its easy detection. In surface display technology, a cell's translocation machinery is harnessed, and a sub-goal of the technology described in Paper II was for our method to allow systematic study of precisely those molecular mechanisms.

3.2 The different compartments of an E. coli cell and how to get there

Around 40% of *E. coli* peptides are moved from the cytosol where they are made, and either inserted into the inner membrane, transported to the periplasm, brought to the outer membrane, or secreted [95]. There are several sophisticated machineries for enabling the translocation of proteins across the different barriers, some better understood than others. The β-barrel assembly machinery (BAM), responsible for insertion of outer membrane proteins into the outer membrane, is still to a large extent a mystery.

For an outer membrane protein (OMP), the journey across the cell wall starts with an N-terminal signal peptide directing the protein to its translocation machinery, continues via one of the translocation pathways across the inner membrane, followed by chaperone-assisted transport through the periplasm, and finally finishes with BAM-assisted outer membrane insertion. An overview of the translocation pathways is presented in Figure 8.

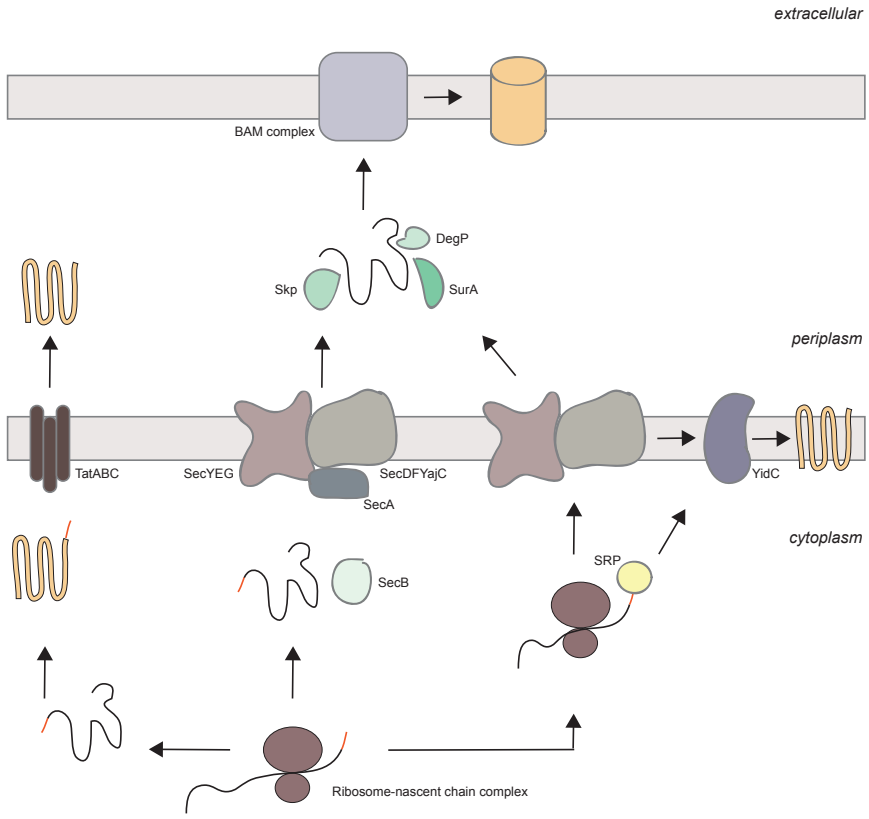


Figure 8. Translocation paths in *E. coli*. Transport of a protein to one of the membranes utilises four major systems. Folded protein translocation is achieved by the Tat (Twin-arginine translocation) pathway. Unfolded proteins are transported across the IM by the SecYEG complex assisted by SecDFYajC. SecB sometimes assist in the transportation of protein to the IM and SecA provides energy for the process via its ATPase activity. The SRP (signal recognition particle) pathway transports proteins that are still being translated to SecYEG, where the protein is translocated using the energy from translation. Inner membrane proteins (IMPs) are moved into the IM by the action of YidC, either together with or without SecYEG. Chaperones Skp, SurA and DegP are important for the transport of outer membrane proteins (OMPs) across the periplasm and to the BAM complex, which inserts OMPs into the OM. Adapted from [96].

3.2.1 Translocation across the inner membrane

Transport across, and insertion of inner membrane proteins (IMPs) into, the inner membrane uses one of three major machineries: the twin-arginine translocation (Tat) pathway for fully folded proteins, the SecYEG complex for unfolded substrates and IMP insertion, or the YidC insertase, which inserts IMPs into the IM alone or together with SecYEG.

Proteins that are to be moved to the IM or beyond contain an N-terminal signal peptide that directs them to their respective translocation pathway. The signal sequence is later cleaved by periplasmic peptidases [97]. The Tat pathway only transports fully folded proteins through the TatABC translocase [98]. These are typically co-factor-dependent proteins, since co-factors can mainly be supplied in the cytoplasm [99].

Unfolded peptides are targeted to the SecYEG translocon, responsible for translocation of around 90% of secreted proteins, either post- or co-translationally. Post-translational targeting to SecYEG is sometimes mediated by the SecB chaperone, preventing folding of the petide while it is transported to SecYEG, but can also be SecB-independent. When the protein has reached the Sec translocon, it is pushed through using energy from the SecA ATPase [100]. The SRP (signal recognition particle) pathway is responsible for co-translational targeting of peptides to the Sec translocon. After binding to the N-terminal signal sequence of a ribosome-nascent chain complex (RNC), SRP is recognised by its receptor FtsY and the RNC directed to the IM and the Sec translocase. The peptide chain is moved through the Sec translocon in a way that is thought to utilise the energy from protein synthesis rather than from SecA [100]. Inner membrane proteins are most often targeted to the SRP pathway [99]

but after docking of the RNC to SecYEG, YidC mediates insertion of the protein into the inner membrane [96]. The opening of a lateral gate in SecYEG enables this movement [99]. Certain inner membrane proteins can be inserted only with YidC, without SecYEG mediation [96]. Additional proteins are important but in somehow unclear ways: SecDFYajC seems to have a role in SecYEG-mediated translocation, and YidD has been shown to be a partner for YidC [96].

3.2.2 Transport through the periplasm

When a protein has reached the periplasm, the N-terminal signal sequence is cleaved off, and the protein either resides in this intermembranic space, or is transported to the outer membrane. The periplasmically located mechanisms are much less well-understood than their cytoplasmic and inner membrane counterparts, but chaperones are essential for delivering outer membrane proteins to the BAM complex [101].

Three chaperones have been identified as key players in transport through the periplasm of *E. coli*: Survival protein A (SurA), Seventeen kilodalton protein (Skp) and DegP [102–104]. SurA is the only one whose absence causes a reduction in the number of outer membrane proteins, and it has been suggested that SurA is the main chaperone used by outer membrane proteins, while they make use of Skp and DegP when necessary [104]. Remarkably, there is no ATP-based or electrochemical energy available in the periplasm, meaning that all actions carried out by chaperones and later by the BAM complex have to do without [104]. The roles of the chaperones are only partly known, but SurA and Skp have been shown to have general chaperone activity, meaning that they bind to unfolded, but not to folded, proteins. SurA also interacts directly with the BAM complex. DegP studies have reported contradicting results, showing both

general chaperone activity, and protease activity for degradation of misfolded or aggregated proteins [105, 106]. Skp and SurA deliver the protein to BAM, which is thought to recognize the so-called beta-motif on the C-terminal strand of OMPs as a signal for outer membrane insertion [107].

3.2.3 Outer membrane insertion

The final step of outer membrane protein biogenesis, the actual insertion of the protein into the outer membrane, is yet more enigmatic than the periplasmic chaperone mechanisms. The BAM complex is required for the process, and its only OMP component, BamA, is one of just two essential OMPs in *E. coli* [104, 108]. In addition to BamA, the BAM complex consists of the four lipoproteins BamB, BamC, BamD, and BamE. BamA forms a beta-barrel structure in the membrane, while BamB-E are associated with the membrane via an N-terminal lipid modification [109, 110]. BamA and BamD are the only essential components of the complex. Two models exist for the insertion of an OMP into the membrane by the BAM complex: the BamA-assisted model and the BamA-budding model. The prior is supported by most scientific evidence, and suggests that the BAM complex primes the membrane by thinning it and destabilising lipids. When an OMP is then delivered by SurA or Skp, it can spontaneously integrate into the membrane. The BamA-budding model instead proposes that the BAM complex pulls the OMP through its central BamA barrel domain, and then lets it out into the membrane via a lateral gate, similar to the SecYEG-YidC-based insertion of IMPs [104].

Independently of the exact details of the mechanisms described in this section, it is clear that the transport of a protein from the cytoplasm all the way to the outer membrane is a very complex

process. Using the described translocation machineries for synthetic biology applications involves big challenges, as well as big opportunities to learn more about the mechanistic details behind. Analogous to how the development of a GFP platform greatly facilitated membrane protein production and studies [111–114], simple assays for the study of surface display and associated translocation mechanisms are needed.

3.3 Major OM anchors and their structure

When it comes to exploitation of the cell machinery for surface display of a protein of choice, many different systems and organisms have been explored. Two of the main anchor types used for bacterial surface display, and the ones chosen in Paper II of this thesis, are autotransporters (ATs) and Outer membrane protein A (OmpA).

3.3.1 Autotransporters

Autotransporters constitute the Type V secretion system, which is commonly found in Gram-negative bacteria. Autotransporters typically function as virulence factors with cytotoxic, adhesive, or proteolytic effects [115]. They also have roles in biofilm formation and immune system evasion. The classic structure of an autotransporter protein is an N-terminal signal peptide, followed by a passenger domain that will be secreted to the outside of the outer membrane, and lastly a C-terminal translocation unit (Figure 9).

Autotransporters are translocated across the inner membrane by means of the SecYEG complex, via the SecA-dependent pathway. Both the SurA and Skp chaperones have been shown to assist the transport across the periplasm [116]. Then, with help from the BAM complex, the C-terminus of the autotransporter is first assembled into

the outer membrane, after which the passenger domain is pulled through the translocation unit.

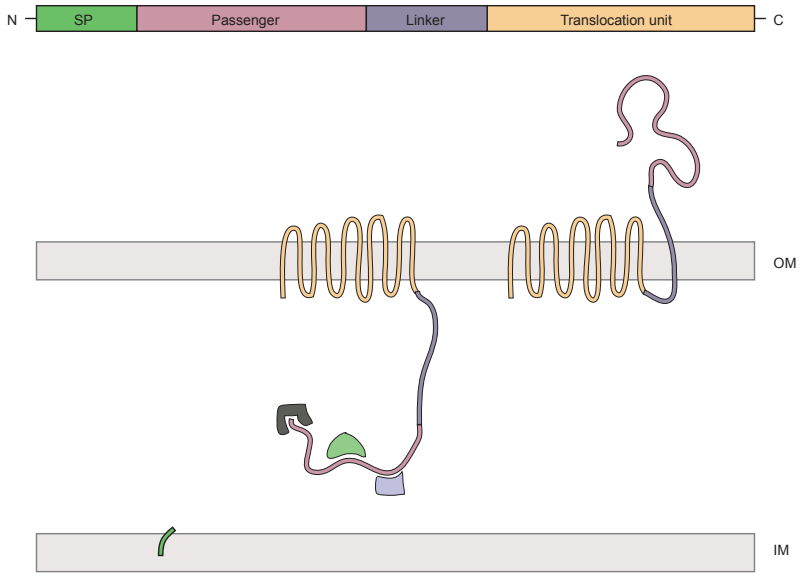


Figure 9. **Autotransporter structure.** Autotransporter proteins consist of an N-terminal signal peptide (SP) for direction to the translocation machinery, a C-terminal translocation unit, and a passenger that is N-terminally linked to the translocation unit. The signal peptide is cleaved off after translocation of the autotransporter into the periplasm. Chaperones assist in transport across the periplasm and the passenger protein is moved to the surface of the cell after insertion of the translocation unit into the OM.

When on the outside of the cell, the passenger is either cleaved, attaches to another site on the outer membrane, or remains bound to the translocation unit [117]. The synthetic biology field has explored autotransporters for the use of bringing various passengers to the cell surface; examples include lipases, dehydrogenases and a bacterial cytochrome P450 enzyme [118–120].

3.3.2 Outer membrane protein A

Outer membrane protein A is a major outer membrane protein in *E. coli*, existing in some 100,000 copies per cell [121]. It forms a beta-barrel with eight transmembrane segments. OmpA serves a multitude of functions, including general porin activity and structure stabilisation (Figure 10) [122, 123]. In addition, it is the receptor for bacterocins and bacteriophages, while also acting in adhesion and invasion itself [123, 124].

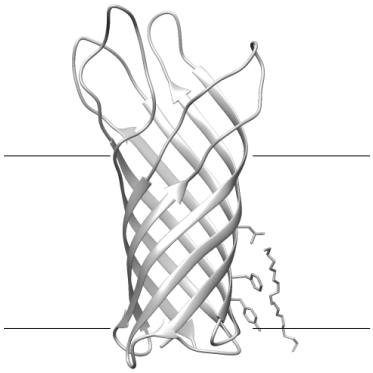


Figure 10. Outer membrane protein A. Structure of OmpA (PDB ID 1bxw). Image created with the UCSF Chimera Package [125].

Many studies have been made on the use of OmpA in synthetic biology, not the least on the fusion version of the protein known as LppOmpA. This hybrid protein consists of the signal peptide of the lipoprotein Lpp, fused to five transmembrane segments of OmpA [126]. It is translocated across the inner membrane using the Sec machinery and the SRP-dependent pathway. Interestingly, YidC also seems to play a role [127]. The LppOmpA anchor has been extensively explored for surface display by C-terminal fusion of a passenger to the hybrid protein, and examples in *E. coli* include

display of hydrolases, lipases and single-chain variable fragments (scFvs) [128–130].

3.4 Harvesting nature's designs: nanobodies and green fluorescent protein

In the development of new synbio technologies, there is an astonishing potential to find suitable solutions among those that nature has already evolved. In Paper II of this thesis, we make use of two outstanding molecules found in nature, and combined with the outer membrane protein machinery of *E. coli* they form the basis for construction of a new surface display platform. In the search for an easy way to detect surface display, it is tempting to use GFP as a reporter [131]. This, however, is a dubious strategy due to the difficulty to discriminate between intracellularly and surface localised GFP. In Paper II of this thesis, we solve this problem by combining GFP with a nanobody (NB) in a new strategy for surface display detection.

3.4.1 Green fluorescent protein

Deep down in the seas, many creatures look exotic from the perspective of us who walk above ground, and transparent organisms with luminous parts are mainstream. In these waters, the *Aequorea victoria* jellyfish swims, glowing intensely green. This caught the attention of Japanese researchers in the 1960s, and led to the nowadays very frequent usage of GFP in cell biology and biotechnology. The history is slightly brutal, with the later Nobel prize-awarded Osamu Shimomura for 19 years spending the summers catching vast numbers of jellyfish (850,000 in the end) from which the fluorescing protein was cut out [132].

Many years after Osamu Shimomura's isolation of the protein, Martin Chalfie, another Nobel laureate, in 1994 succeeded in cloning and expressing the GFP-encoding gene in both *E. coli* and the nematode worm *C. elegans* [133]. This was the start of a GFP revolution, and we have since seen both fluorescing cats and mice [134]. In biotechnology, GFP is typically used as a reporter, showing for example when and where a gene is expressed, and it has also been instrumental for optimisation of membrane protein production [114]. GFP has been used to study many fundamentally important processes in biology such as cell division and inner membrane protein topology [111, 135]. Precisely the localisation aspect was what we used in the study in Paper II of this thesis. GFP is a useful tool not only because of its fluorescence, but also because it is a highly stable protein that is easy to produce and work with in the laboratory. Its value is in essence that it enables the visualisation of the normally invisible, a capacity that anyone working with microorganisms, proteins and genes can appreciate (Figure 11).

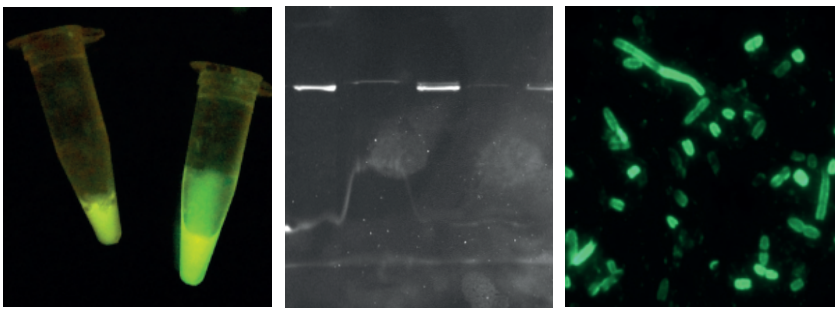


Figure 11. Green fluorescent protein in the lab. GFP enables visualisation of protein **(A)** in cell culture, **(B)** on protein gels, and **(C)** in microscopy.

3.4.2 Nanobodies

Another example of remarkable molecules are the tiny single-chain antibodies found in sharks as well as camelid species (camels,

dromedaries, llamas), which due to their relatively small size are called nanobodies. These heavy-chain-only antibody molecules were first reported in camel serum in 1993 and subsequent characterisation of NBs has shown their potential for use as both therapeutic agents and biotechnological alternatives to conventional antibodies [136]. Nanobodies are around 12 kDa in size and carry out the normal antibody functions, in spite of being just a fraction of the size of a conventional antibody (Figure 12).

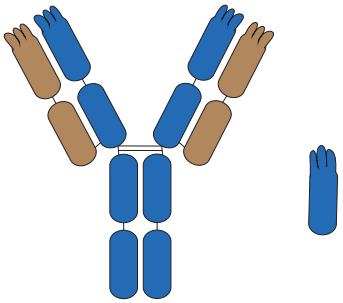


Figure 12. Antibody versus nanobody. Conventional antibodies (left) consist of two heavy chains (blue) and two light chains (orange), linked by disulphide bonds. Nanobodies (right) consist of only one heavy variable domain.

NBs bind with high specificity and affinity to their antigens, after which they recruit the immune response apparatus in order to attack the foreign molecule. Their epitopes differ from those of ordinary antibodies in that NBs are more often targeting concave epitopes, typically found in the active site of enzymes [137]. Hence, NBs also often modify the activity of their protein antigen [137–139]. Their antigen binding affinity is in the nanomolar range, comparable to high-affinity antibodies and similar to that of scFvs [140]. Furthermore, they are highly soluble, very stable, and easily produced in heterologous hosts [140, 141]. The desired binder can be

selected via phage display but nature's own method, with prior immunisation of a camel, is recommended to achieve higher affinity NBs [140, 142, 143]. The many advantageous properties of NBs have sparked their use for therapeutic methods as well as biotechnological purposes [144, 145].

3.4.3 GFP and GFP-NBs

In 2010, Kirchhofer *et al*. injected a camel with GFP, collected its nanobody repertoire from the B-lymphocytes, and cloned it into a phagemid library subsequently used to select GFP-binders [143]. Specifically, they isolated an enhancer-NB, increasing GFP's fluorescence 10-fold, and a minimiser-NB causing GFP's fluorescence to decrease by a factor five. Together, these findings resulted in a new toolbox consisting of GFP and two nanobodies, capable of modifying GFP's fluorescence properties. The affinity of NB to GFP is very high, and remarkably, the interaction is highly stable. As described in Paper II, this enables analysis of a GFP-NB complex with many different methods, such as denaturing protein gels (Figure 13).

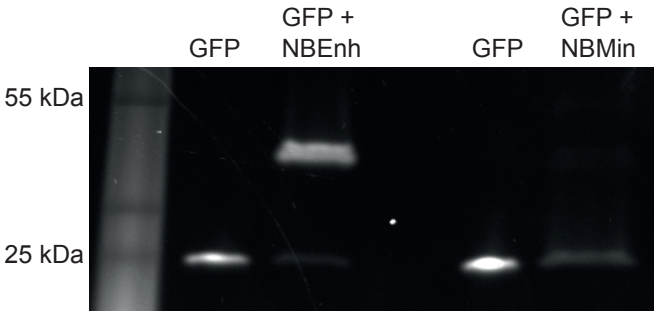


Figure 13. GFP-nanobodies' effect on GFP. In-gel fluorescence shows how GFP binds to both enhancer-NB, resulting in a fluorescent complex, and to minimiser-NB, resulting in a marked decrease in fluorescence.

In Paper II of this thesis, we exploited the GFP:nanobody interaction for development of a new platform for detection of surface displayed protein. The enhancer-nanobody is in this method fused to a surface anchored enzyme and subsequently detected and analysed using free GFP. In this way, knowledge from basic biological research is used to adjust a living cell according to our needs, while in a double sense making use of the clever systems already developed by nature.

Interestingly, we observed large differences in display efficiency between the OmpA anchor and the autotransporter anchor C-IgAP, and also a cargo-dependency. We hypothesise that this is partly due to differences in the translocation paths taken, which in turn is dependent on the choice of anchor and signal peptide. This hypothesis is supported by a test of a small library of signal peptides' effect on surface display, showing a clear impact for especially the autotransporter anchor. The library contained signal peptides for direction to the Tat, SecB and SRP-dependent translocation pathways, and our comparison showed an anticipated disadvantage of using the Tat pathway. The experiment highlighted how balancing of the different components involved is crucial for successful display; illustrating this is the observation that the signal peptide giving the highest display when produced from a low-copy plasmid was no longer beneficial in a different backbone.

In general, we observed difficulties for the cell to produce the autotransporter C-IgAP. The fact that OmpA is an *E. coli* protein whereas C-IgAP is not might be part of the explanation. Additionally, the observations could potentially be due to saturation of the translocation machinery, an imbalance that is presumably relieved by adding the enzyme Chitinase A to the protein fusion. One can speculate that the autotransporter protein sequence, with the

passenger always being N-terminal to the translocation domain, leads to the passenger having an increased influence on translocation efficiency.

Conclusion and perspectives

Synthetic biology offers the opportunity to explore and make use of the clever solutions existing in the natural world around us. This thesis is about the two pillars of synthetic biology, and how they can be used to understand and (re)construct living systems. The work underlying the thesis spans from increasing our understanding of how *Escherichia coli* adapts to a challenging environment, to using characterised parts of the same organism for the clear biotechnological purpose of surface display.

Paper I describes the study of adaptive mutation formation in ageing *E. coli* cells. It reveals the occurrence of non-random mutagenesis patterns when cells are starved for an extended time. This paper casts new light on the complex mechanisms behind adaptive mutagenesis, a phenomenon that nevertheless remains occult. Many follow-up experiments are presently on-going in our lab, investigating the importance of codon usage for the mutability of genes, deepening our understanding of the timeliness of different mutation types, and exploring the possibility of designing a protein for mutation and evolution.

In Paper II, mechanisms that to a large extent have been deciphered are used for the development of a new surface display platform. By combining GFP and nanobodies, we address the need for an easy detection system. In parallel, we used the new assay for evaluation of signal peptide influence on display efficiency, and also observed large differences in display levels depending on the anchors and

passengers used. A future prospect is for the surface display platform to be a robust tool for applied as well as fundamental studies, linking application with basic science. Our hope is that the platform can extend to the basics of synbio and be used to deepen the current knowledge of translocation mechanisms such as those associated with the still mysterious BAM complex. For biotechnological developments, future uses should include strain optimisation and screening. For example one could envision screening mutant libraries for better surface display with our approach.

The study of mutation generation in a whole-genome context would not have been possible without the recent advances in sequencing, with Next generation sequencing (NGS) becoming affordable and thereby accessible at a new scale. Rapid decreases in prices for important technologies have indeed been instrumental for their widespread use throughout the synbio community, all the way to the new citizen scientists and do-it-yourself-biologists. Similar to the development of computer science and technology, democratising synbio technology might be a recipe for success and scientific progress. From that perspective, a particularly positive feature of the NB:GFP platform is that GFP is easily and cheaply produced in *E. coli*, making the assay possible to perform in any lab. In the open-source spirit of synbio, our assay makes surface display analysis more accessible, facilitating developments of this technology as well as studies of the underlying translocation mechanisms.

This thesis illustrates the two synbio pillars of understanding and constructing organisms, and how the two overlap and complement each other. By pursuing an open-minded strategy combining basic and applied research, synbio constitutes a chance to create the

sustainable solutions we are in great need of. The work presented in this thesis will hopefully be a fruitful contribution to this quest.

References

1. Hunter P: **Massing life**. *EMBO Rep* 2010, **11**:511–4.

2. Rothschild LJ, Mancinelli RL: **Life in extreme environments**. *Life Extrem Environ* 2007, **409**(September 2000):1092–1101.

3. Coil DA, Neches RY, Lang JM, Brown WE, Severance M, Cavalier DD, Eisen JA: **Growth of 48 built environment bacterial isolates on board the International Space Station (ISS)**. *PeerJ* 2016, **4**:1–11.

4. Luckey TD: **Introduction to intestinal microecology**. *Am J Clin Nutr* 1972, **25**:1292–1294.

5. Sender R, Fuchs S, Milo R: **Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans**. *Cell* 2016, **164**:337–340.

6. Linko YY, Javanainen P, Linko S: **Biotechnology of bread baking**. *Trends Food Sci Technol* 1997, **8**:339–344.

7. United Nations: *Resolution Adopted by the General Assembly on 25 September 2015*. 2015:1–35.

8. Ceballos G, Ehrlich PR, Barnosky AD, García A, Pringle RM, Palmer TM: **Accelerated modern human–induced species losses: entering the sixth mass extinction**. *Sci Adv* 2015, **1**:1–5.

9. Pachauri RK, Meyer LA: *IPCC, 2014: Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Geneva, Switzerland: IPCC; 2014.

10. WWF: *Living Blue Planet Report. Species, Habitats and Human Well-Being. Volume 9*. Gland, Switzerland: WWF; 2015.

11. International Energy Agency: **Key World Energy Statistics 2015**. 2015:81.

12. Jordbruksverket: *Köttkonsumtionen i siffror. Utveckling och orsaker*. 2013.

13. Biltoft-Jensen A, Kørup K, Christensen T, Eriksen K, Ygil KH, Fagt S: *Køds Rolle i Kosten*. DTU Fødevareinstituttet; 2016.

14. EU: *Regulation (EC) No 1829/2003 of the European Parliament and of the Council on Genetically Modified Food and Feed*. *Volume L 268*; 2003:1 – 23.

15. Hayden EC: **Tech investors bet on synthetic biology**. *Nature* 2015, **527**:9.

16. Szybalski W: ***In vivo* and *in vitro* initiation of transcription**. In *Control of*

gene expression. *Volume 405*. Edited by Kohn A, Shatkey A. New York: New York Plenum Press; 1974.

17. Church GM, Elowitz MB, Smolke CD, Voigt CA, Weiss R: **Realizing the potential of synthetic biology.** *Nat Rev Mol Cell Biol* 2014, **15**:289–94.

18. IBioEducation: **Synthetic Biology: Overview - Victor de Lorenzo**. *Online video* 2016:Accessed 19 April 2016.

19. VitruvianMan07: **iGEM - Drew Endy Defining Synthetic Biology**. *Online video* 2007:Accessed 19 April 2016.

20. Science and Technology Select Committee: *Genetically Modified Insects*. London, UK; 2015(December).

21. Cameron DE, Bashor CJ, Collins JJ: **A brief history of synthetic biology.** *Nat Rev Microbiol* 2014, **12**:381–90.

22. Minssen T, Wested JB: **Standardization, IPRs and Open Innovation in Synthetic Biology**. *Innov Compet Collab* 2014:1–20.

23. Delgado A, Porcar M: **Designing *de novo*: interdisciplinary debates in synthetic biology.** *Syst Synth Biol* 2013, **7**:41–50.

24. Schmidt M, Ganguli-Mitra A, Torgersen H, Kelle A, Deplazes A, Biller-Andorno N: **A priority paper for the societal and ethical aspects of synthetic biology**. *Syst Synth Biol* 2009, **3**:3–7.

25. Smolke CD, Silver PA: **Informing biological design by integration of systems and synthetic biology**. *Cell* 2011, **144**:855–859.

26. Pardee K, Green AA, Ferrante T, Cameron DE, Daleykeyser A, Yin P, Collins JJ: **Paper-based synthetic gene networks**. *Cell* 2014, **159**:940–954.

27. Hocking MB: **Vanillin: Synthetic Flavoring from Spent Sulfite Liquor**. *J Chem Educ* 1997, **74**:1055.

28. Walsh G: **Biopharmaceutical benchmarks 2014**. *Nat Biotechnol* 2014, **32**:992–1000.

29. Schmidt FR: **Recombinant expression systems in the pharmaceutical industry**. *Appl Microbiol Biotechnol* 2004, **65**:363–372.

30. Terpe K: **Overview of bacterial expression systems for heterologous protein production: From molecular and biochemical fundamentals to commercial systems**. *Appl Microbiol Biotechnol* 2006, **72**:211–222.

31. Lee SY, Mattanovich D, Villaverde A: **Systems metabolic engineering, industrial biotechnology and microbial cell factories.** *Microb Cell Fact* 2012,

11:156.

32. Daegelen P, Studier FW, Lenski RE, Cure S, Kim JF: **Tracing Ancestors and Relatives of *Escherichia coli* B, and the Derivation of B Strains REL606 and BL21(DE3)**. *J Mol Biol* 2009, **394**:634–643.

33. Sezonov G, Joseleau-Petit D, D'Ari R: ***Escherichia coli* physiology in Luria-Bertani broth**. *J Bacteriol* 2007, **189**:8746–8749.

34. Rosano GL, Ceccarelli EA: **Recombinant protein expression in *Escherichia coli*: advances and challenges.** *Front Microbiol* 2014, **5**(April):172.

35. Lederberg J, Tatum EL: **Gene Recombination in *Escherichia coli***. *Nature* 1946, **158**:558.

36. Jacob F, Monod J: **Genetic regulatory mechanisms in the synthesis of proteins.** *J Mol Biol* 1961, **3**:318–356.

37. Benzer S: **On the topography of the genetic fine structure**. *Proc Natl Acad Sci* 1961, **47**:403–415.

38. Zimmer C: *Microcosm - E. Coli and the New Science of Life*. United States: Vintage Books; 2008.

39. Chen X, Zhou L, Tian K, Kumar A, Singh S, Prior BA, Wang Z: **Metabolic engineering of *Escherichia coli*: A sustainable industrial platform for bio-based chemical production**. *Biotechnol Adv* 2013, **31**:1200–1223.

40. **Human insulin receives FDA approval**. *FDA Drug Bull* 1982, **12**:18–19.

41. Blattner FR, Plunkett 3rd G, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y: **The complete genome sequence of *Escherichia coli* K-12**. *Science* 1997, **277**:1453–1462.

42. Jeong H, Barbe V, Lee CH, Vallenet D, Yu DS, Choi SH, Couloux A, Lee SW, Yoon SH, Cattolico L, Hur CG, Park HS, Ségurens B, Kim SC, Oh TK, Lenski RE, Studier FW, Daegelen P, Kim JF: **Genome Sequences of *Escherichia coli* B strains REL606 and BL21(DE3)**. *J Mol Biol* 2009, **394**:644–652.

43. Lamarck J-B: *Philosophie Zoologique, Ou Exposition Des Considérations Relatives À L'histoire Naturelle Des Animaux*. Paris; 1809.

44. Darwin C: *On the Origin of Species*. 1st edition. London; 1859.

45. Gould S: *The Structure of Evolutionary Theory*. Cambridge, MA: Harvard University Press; 2002.

46. Luria SE, Delbrück M: **Mutations of Bacteria from Virus Sensitivity to Virus Resistance.** *Genetics* 1943, **28**:491–511.

47. Masel J: **Genetic drift**. *Curr Biol* 2011, **21**:R837–R838.

48. Koonin E V, Wolf YI: **Is evolution Darwinian or/and Lamarckian?** *Biol Direct* 2009, **4**:42.

49. Foster P: **Stress-induced mutagenesis in bacteria**. *Science* 2007, **300**:1404–1409.

50. Galhardo RS, Hastings PJ, Rosenberg SM: **Mutation as a Stress Response and the Regulation of Evolvability**. *Crit Rev Biochem Mol Biol* 2007, **42**:399–435.

51. Cairns J, Overbaugh J, Miller S: **The origin of mutants**. *Nature* 1988, **335**:142–145.

52. Cairns J, Foster PL: **Adaptive reversion of a frameshift mutation in *E. coli***. *Genetics* 1991, **128**:695–701.

53. Foster PL: **Population dynamics of a Lac- strain of *Escherichia coli* during selection for lactose utilization**. *Genetics* 1994, **138**:253–261.

54. Maisnier-Patin S, Roth JR: **The origin of mutants under selection: How natural selection mimics mutability (adaptive mutation)**. *Microb Evol* 2015.

55. Morreall J, Kim A, Liu Y, Degtyareva N, Weiss B, Doetsch PW: **Evidence for Retromutagenesis as a Mechanism for Adaptive Mutation in *Escherichia coli***. *PLoS Genet* 2015, **11**:1–12.

56. Roth JR, Kugelberg E, Reams AB, Kofoid E, Andersson DI: **Origin of mutations under selection: the adaptive mutation controversy.** *Annu Rev Microbiol* 2006, **60**:477–501.

57. Foster PL, Cairns J: **Mechanisms of directed mutation**. *Genetics* 1992, **131**:783–789.

58. Hall BG: **Spontaneous point mutations that occur more often when advantageous than when neutral**. *Genetics* 1990, **126**:5–16.

59. Torkelson J, Harris RS, Lombardo MJ, Nagendran J, Thulin C, Rosenberg SM: **Genome-wide hypermutation in a subpopulation of stationary-phase cells underlies recombination-dependent adaptive mutation**. *EMBO J* 1997, **16**:3303–3311.

60. Rosche WA, Foster PL: **The role of transient hypermutators in adaptive mutation in *Escherichia coli*.** *Proc Natl Acad Sci U S A* 1999, **96**:6862–7.

61. Susan Slechta E, Liu J, Andersson DI, Roth JR: **Evidence that selected amplification of a bacterial lac frameshift allele stimulates Lac+ reversion (adaptive mutation) with or without general hypermutability**. *Genetics* 2002, **161**:945–956.

62. Roth JR, Kofoid E, Roth FP, Berg OG, Seger J, Andersson DI: **Regulating general mutation rates: Examination of the hypermutable state model for Cairnsian adaptive mutation**. *Genetics* 2003, **163**:1483–1496.

63. Silver S: **Beyond the fringe: When science moves from innovative to nonsense**. *FEMS Microbiol Lett* 2014, **350**:2–8.

64. Davis BD: **Transcriptional bias: a non-Lamarckian mechanism for substrate-induced mutations**. *Proc Natl Acad Sci U S A* 1989, **86**:5005–5009.

65. Bridges BA: **mutY "directs" mutation?** *Nature* 1995, **375**:741.

66. Sekowska A, Wendel S, Nørholm MHH, Danchin A: **Generation of mutation hotspots in ageing bacterial colonies**. *bioRxiv* 2016:http://dx.doi.org/10.1101/041525.

67. You C, Okano H, Hui S, Zhang Z, Kim M, Gunderson CW, Wang YP, Lenz P, Yan D, Hwa T: **Coordination of bacterial proteome with metabolism by cyclic AMP signalling**. *Nature* 2013, **500**:301–306.

68. Kolb A, Busby S, Buc H, Garges S, Adhya S: **Transcriptional regulation by cAMP and its receptors**. *AnnRevBiochem* 1993, **62**:749–795.

69. Körner H, Sofia HJ, Zumft WG: **Phylogeny of the bacterial superfamily of Crp-Fnr transcription regulators: Exploiting the metabolic spectrum by controlling alternative gene programs**. *FEMS Microbiol Rev* 2003, **27**:559–592.

70. Hall BG: **Spectrum of mutations that occur under selective and non-selective conditions in *E. coli***. *Genetica* 1991, **84**:73–76.

71. Bridges BA: **Mutation in resting cells: the role of endogenous DNA damage**. *Cancer Surv* 1996, **28**:155–167.

72. Brégeon D, Doddridge Z a., You HJ, Weiss B, Doetsch PW: **Transcriptional mutagenesis induced by uracil and 8-oxoguanine in *Escherichia coli***. *Mol Cell* 2003, **12**:959–970.

73. Collins DW, Jukes TH: **Rates of transition and transversion in coding sequences since the human-rodent divergence.** *Genomics* 1994:386–396.

74. Tarassova K, Tegova R, Tover A, Teras R, Tark M, Saumaa S, Kivisaar M: **Elevated mutation frequency in surviving populations of carbon-starved rpoS-deficient *Pseudomonas putida* is caused by reduced expression of superoxide dismutase and catalase**. *J Bacteriol* 2009, **191**:3604–3614.

75. Ishizuka H, Hanamura A, Inada T, Aiba H: **Mechanism of the down-regulation of cAMP receptor protein by glucose in *Escherichia coli*: role of autoregulation of the *crp* gene**. *EMBO J* 1994, **13**:3077–3082.

76. Hengge R: **The two-component network and the general stress sigma factor RpoS (sigma S) in *Escherichia coli***. *Adv Exp Med Biol* 2008, **631**:40–53.

77. Franchini AG, Ihssen J, Egli T: **Effect of global regulators RpoS and cyclic-AMP/CRP on the catabolome and transcriptome of *Escherichia coli* K12 during carbon- and energy-limited growth**. *PLoS One* 2015, **10**:1–24.

78. Marraffini LA, Sontheimer EJ: **CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea**. *Nat Rev Genet* 2010, **11**:181–190.

79. Yosef I, Goren MG, Qimron U: **Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli***. *Nucleic Acids Res* 2012, **40**:5569–5576.

80. Koonin E V., Wolf YI: **Just how Lamarckian is CRISPR-Cas immunity: the continuum of evolvability mechanisms**. *Biol Direct* 2016, **11**:9.

81. Koonin E V., Wolf YI: **Evolution of microbes and viruses: a paradigm shift in evolutionary biology?** *Front Cell Infect Microbiol* 2012, **2**:1–15.

82. Saier MH: **Microcompartments and protein machines in prokaryotes.** *J Mol Microbiol Biotechnol* 2013, **23**:243–69.

83. Seppälä S: **Dual-topology membrane proteins in *Escherichia coli***. Stockholm University; 2011.

84. Shibuya I: **Metabolic regulations and biological functions of phospholipids in *Escherichia coli*.** *Prog Lipid Res* 1992, **31**:245–99.

85. Raetz CRH, Whitfield C: **Lipopolysaccharide Endotoxins**. *Annu Rev Biochem* 2002, **71**:1–57.

86. Wallin E, von Heijne G: **Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms.** *Protein Sci* 1998, **7**:1029–38.

87. Overington JP, Al-Lazikani B, Hopkins AL: **How many drug targets are there?** *Nat Rev Drug Discov* 2006, **5**:993–6.

88. Hopkins AL, Groom CR: **The druggable genome.** *Nat Rev Drug Discov* 2002, **1**:727–30.

89. Hutchison CA, Chuang R-Y, Noskov VN, Assad-Garcia N, Deerinck TJ, Ellisman MH, Gill J, Kannan K, Karas BJ, Ma L, Pelletier JF, Qi Z-Q, Richter RA, Strychalski EA, Sun L, Suzuki Y, Tsvetanova B, Wise KS, Smith HO, Glass JI, Merryman C, Gibson DG, Venter JC: **Design and synthesis of a minimal bacterial genome**. *Science* 2016, **351**:aad6523.

90. Smith GP: **Filamentous Fusion Phage: Novel Expression Vectors that Display Cloned Antigens on the Virion Surface**. *Science* 1985, **228**:1315–1317.

91. Wu C-H, Liu I-J, Lu R-M, Wu H-C: **Advancement and applications of peptide phage display technology in biomedical science**. *J Biomed Sci* 2016, **23**:8.

92. Volk A, Hu FJ, Rockberg J: **Epitope Mapping of Monoclonal and Polyclonal Antibodies Using Bacterial Cell Surface Display**. *Methods Mol Biol* , **1131**:485–500.

93. Lee SY, Choi JH, Xu Z: **Microbial cell-surface display**. *Trends Biotechnol* 2003, **21**:45–52.

94. Schüürmann J, Quehl P, Festel G, Jose J: **Bacterial whole-cell biocatalysts by surface display of enzymes: toward industrial application.** *Appl Microbiol Biotechnol* 2014, **98**:8031–46.

95. Weiner JH, Li L: **Proteome of the *Escherichia coli* envelope and technological challenges in membrane proteome analysis**. *Biochim Biophys Acta - Biomembr* 2008, **1778**:1698–1713.

96. Boock JT, Waraho-Zhmayev D, Mizrachi D, Delisa MP: *Beyond the Cytoplasm of Escherichia Coli: Localizing Recombinant Proteins Where You Want Them*. *Volume 1258*. New York, NY: Springer New York; 2015. [*Methods in Molecular Biology*]

97. Hegde RS, Bernstein HD: **The surprising complexity of signal sequences**. *Trends Biochem Sci* 2006, **31**:563–571.

98. Palmer T, Berks BC: **The twin-arginine translocation (Tat) protein export pathway.** *Nat Rev Microbiol* 2012, **10**:483–96.

99. Kudva R, Denks K, Kuhn P, Vogt A, Müller M, Koch H-G: **Protein translocation across the inner membrane of Gram-negative bacteria: the Sec and Tat dependent protein transport pathways.** *Res Microbiol* 2013,

164:505–34.

100. Beckwith J: **The Sec-dependent pathway.** *Res Microbiol* 2013, **164**:497–504.

101. McMorran LM, Brockwell DJ, Radford SE: **Mechanistic studies of the biogenesis and folding of outer membrane proteins in vitro and in vivo: What have we learned to date?** *Arch Biochem Biophys* 2014, **564**:265–280.

102. Behrens S, Maier R, De Cock H, Schmid FX, Gross CA: **The SurA periplasmic PPIase lacking its parvulin domains functions in vivo and has chaperone activity**. *EMBO J* 2001, **20**:285–294.

103. Sklar JG, Wu T, Kahne D, Silhavy TJ: **Defining the roles of the periplasmic chaperones SurA, Skp, and DegP in *Escherichia coli***. *Genes Dev* 2007, **21**:2473–2484.

104. Rollauer SE, Sooreshjani MA, Noinaj N, Buchanan SK, Buchanan SK: **Outer membrane protein biogenesis in Gram-negative bacteria**. 2015.

105. Spiess C, Beil A, Ehrmann M: **A temperature-dependent switch from chaperone to protease in a widely conserved heat shock protein**. *Cell* 1999, **97**:339–347.

106. Ge X, Wang R, Ma J, Liu Y, Ezemaduka AN, Chen PR, Fu X, Chang Z: **DegP primarily functions as a protease for the biogenesis of β-barrel outer membrane proteins in the Gram-negative bacterium *Escherichia coli***. *FEBS J* 2014, **281**:1226–1240.

107. Robert V, Volokhina EB, Senf F, Bos MP, Van Gelder P, Tommassen J: **Assembly factor Omp85 recognizes its outer membrane protein substrates by a species-specific C-terminal motif**. *PLoS Biol* 2006, **4**:1984–1995.

108. Wu T, Malinverni J, Ruiz N, Kim S, Silhavy TJ, Kahne D: **Identification of a multicomponent complex required for outer membrane biogenesis in *Escherichia coli***. *Cell* 2005, **121**:235–245.

109. Albrecht R, Schütz M, Oberhettinger P, Faulstich M, Bermejo I, Rudel T, Diederichs K, Zeth K: **Structure of BamA, an essential factor in outer membrane protein biogenesis**. *Acta Crystallogr Sect D Biol Crystallogr* 2014, **70**:1779–1789.

110. Hagan CL, Silhavy TJ, Kahne D: **β-Barrel Membrane Protein Assembly by the Bam Complex**. *Annu Rev Biochem* 2011, **80**:189–210.

111. Daley DO, Rapp M, Granseth E, Melén K, Drew D, von Heijne G: **Global topology analysis of the *Escherichia coli* inner membrane**

proteome. *Science* 2005, **308**:1321–3.

112. Sonoda Y, Newstead S, Hu N-J, Alguel Y, Nji E, Beis K, Yashiro S, Lee C, Leung J, Cameron AD, Byrne B, Iwata S, Drew D: **Benchmarking membrane protein detergent stability for improving throughput of high-resolution X-ray structures.** *Structure* 2011, **19**:17–25.

113. Drew D, Lerch M, Kunji E, Slotboom D-J, de Gier J-W: **Optimization of membrane protein overexpression and purification using GFP fusions.** *Nat Methods* 2006, **3**:303–313.

114. Drew DE, von Heijne G, Nordlund P, de Gier JW: **Green fluorescent protein as an indicator to monitor membrane protein overexpression in *Escherichia coli*.** *FEBS Lett* 2001, **507**:220–4.

115. van Ulsen P, Rahman SU, Jong WSP, Daleke-Schermerhorn MH, Luirink J: **Type V secretion: from biogenesis to biotechnology.** *Biochim Biophys Acta* 2014, **1843**:1592–611.

116. Leyton DL, Rossiter AE, Henderson IR: **From self sufficiency to dependence: mechanisms and factors important for autotransporter biogenesis.** *Nat Rev Microbiol* 2012, **10**:213–25.

117. Henderson IR, Navarro-Garcia F, Desvaux M, Fernandez RC, Ala'Aldeen D: **Type V Protein Secretion Pathway : the Autotransporter Story**. *Microbiol Mol Biol Rev* 2004, **68**:692–744.

118. Schumacher SD, Hannemann F, Teese MG, Bernhardt R, Jose J: **Autodisplay of functional CYP106A2 in *Escherichia coli*.** *J Biotechnol* 2012, **161**:104–12.

119. Kranen E, Detzel C, Weber T, Jose J: **Autodisplay for the co-expression of lipase and foldase on the surface of *E. coli*: washing with designer bugs.** *Microb Cell Fact* 2014, **13**:19.

120. Jose J, von Schwichow S: **Autodisplay of active sorbitol dehydrogenase (SDH) yields a whole cell biocatalyst for the synthesis of rare sugars.** *Chembiochem* 2004, **5**:491–9.

121. Koebnik R, Locher KP, Van Gelder P: **Structure and function of bacterial outer membrane proteins: barrels in a nutshell.** *Mol Microbiol* 2000, **37**:239–53.

122. Sugawara E, Nikaido H: **Pore-forming activity of OmpA protein of *Escherichia coli***. *J Biol Chem* 1992, **267**:2507–2511.

123. Krishnan S, Prasadarao N V: **Outer membrane protein A and OprF –**

Versatile roles in Gram-negative bacterial infections. *FEBS J* 2012, **279**:919–931.

124. Smith SGJ, Mahon V, Lambert MA, Fagan RP: **A molecular Swiss army knife: OmpA structure, function and expression**. *FEMS Microbiol Lett* 2007, **273**:1–11.

125. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE: **UCSF Chimera - A visualization system for exploratory research and analysis**. *J Comput Chem* 2004, **25**:1605–1612.

126. Earhart CF: **Use of an Lpp-OmpA Fusion Vehicle for Bacterial Surface Display**. 2000, **326**:506–516.

127. Fröderberg L, Houben ENG, Baars L, Luirink J, De Gier JW: **Targeting and translocation of two lipoproteins in *Escherichia coli* via the SRP/Sec/YidC pathway**. *J Biol Chem* 2004, **279**:31026–31032.

128. Liu R, Yang C, Xu Y, Xu P, Jiang H, Qiao C: **Development of a whole-cell biocatalyst/biosensor by display of multiple heterologous proteins on the *Escherichia coli* cell surface for the detoxification and detection of organophosphates**. *J Agric Food Chem* 2013, **61**:7810–7816.

129. Daugherty PS, Chen G, Olsen MJ, Iverson BL, Georgiou G: **Antibody affinity maturation using bacterial surface display.** *Protein Eng* 1998, **11**:825–832.

130. Jo J, Han C, Kim S, Kwon H, Lee H: **Surface Display Expression of *Bacillus licheniformis* Lipase in *Escherichia coli* Using Lpp'OmpA Chimera**. *J Microbiol* 2014, **52**.

131. Sun F, Pang X, Xie T, Zhai Y, Wang G, Sun F: **BrkAutoDisplay: functional display of multiple exogenous proteins on the surface of *Escherichia coli* by using BrkA autotransporter**. *Microb Cell Fact* 2015, **14**:129.

132. Nobel.org: **The Nobel Prize in Chemistry 2008: "for the discovery and development of the green fluorescent protein, GFP."** *Nobelprize.org* 2008, **38**:2821–2.

133. Chalfie M, Tu Y, Euskirchen G, Ward WW, Prasher DC: **Green fluorescent protein as a marker for gene expression**. *Science* 1994, **263**:802–805.

134. Wongsrikeao P, Saenz D, Rinkoski T, Otoi T, Poeschla E: **Antiviral restriction factor transgenesis in the domestic cat**. *Nat Methods* 2011, **8**:853–

859.

135. Skoog K, Söderström B, Widengren J, Von Heijne G, Daley DO: **Sequential closure of the cytoplasm and then the periplasm during cell division in *Escherichia coli***. *J Bacteriol* 2012, **194**:584–586.

136. Hamers-Casterman C, Atarhouch T, Muyldermans S, Robinson G, Hamers C, Songa EB, Bendahman N, Hamers R: **Naturally occurring antibodies devoid of light chains.** *Nature* 1993, **363**:446–8.

137. Lauwereys M, Ghahroudi MA, Desmyter A, Kinne J, Hölzer W, De Genst E, Wyns L, Muyldermans S: **Potent enzyme inhibitors derived from dromedary heavy-chain antibodies**. *EMBO J* 1998, **17**:3512–3520.

138. Desmyter A, Spinelli S, Payan F, Lauwereys M, Wyns L, Muyldermans S, Cambillau C: **Three camelid VHH domains in complex with porcine pancreatic alpha-amylase: Inhibition and versatility of binding topology**. *J Biol Chem* 2002, **277**:23645–23650.

139. Saerens D, Kinne J, Bosmans E, Wernery U, Muyldermans S, Conrath K: **Single domain antibodies derived from dromedary lymph node and peripheral blood lymphocytes sensing conformational variants of prostate-specific antigen**. *J Biol Chem* 2004, **279**:51965–51972.

140. Muyldermans S, Baral TN, Retamozzo VC, De Baetselier P, De Genst E, Kinne J, Leonhardt H, Magez S, Nguyen VK, Revets H, Rothbauer U, Stijlemans B, Tillib S, Wernery U, Wyns L, Hassanzadeh-Ghassabeh G, Saerens D: **Camelid immunoglobulins and nanobody technology**. *Vet Immunol Immunopathol* 2009, **128**:178–183.

141. Olichon A, Surrey T: **Selection of Genetically Encoded Fluorescent Single Domain Antibodies Engineered for Efficient Expression in *Escherichia coli***. *J Biol Chem* 2007, **282**:36314–36320.

142. Tanha J, Xu P, Chen Z, Ni F, Kaplan H, Narang SA, MacKenzie CR: **Optimal Design Features of Camelized Human Single-domain Antibody Libraries**. *J Biol Chem* 2001, **276**:24774–24780.

143. Kirchhofer A, Helma J, Schmidthals K, Frauer C, Cui S, Karcher A, Pellis M, Muyldermans S, Casas-delucchi CS, Cardoso MC, Leonhardt H, Hopfner K, Rothbauer U: **Modulation of protein properties in living cells using nanobodies**. *Nat Struct Mol Biol* 2009, **17**:133–138.

144. Steeland S, Vandenbroucke RE, Libert C: **Nanobodies as therapeutics: big opportunities for small antibodies**. *Drug Discov Today* 2016, **00**.

145. Schmidthals K, Helma J, Zolghadr K, Rothbauer U, Leonhardt H: **Novel antibody derivatives for proteome and high-content analysis.** *Anal Bioanal Chem* 2010, **397**:3203–8.

Paper I

Generation of mutation hotspots in ageing bacterial colonies

Agnieszka Sekowska, Sofie Wendel, Morten Nørholm, Antoine Danchin

Abstract

How do ageing bacterial colonies generate adaptive mutants? Over a period of two months, we isolated on ageing colonies outgrowing mutants able to use a new carbon source, and sequenced their genomes. This allowed us to uncover exquisite details on the molecular mechanism behind their adaptation: most mutations were located in just a few hotspots in the genome and over time, mutations increasingly originated from 8-oxo-guanosine, formed exclusively on the transcribed strand.

Introduction

Bacteria constitute a precious biological model system for studying the molecular details of ageing and evolution. Bacterial cells defective in the MutY enzyme, responsible for removing adenine nucleotides pairing with the 8-oxo oxidized variant of guanosine (8-oxo-G), exhibit a dramatic increase in the number of adaptive mutants and Bridges proposed a model to explain this phenomenon that was later termed retromutagenesis [1,2]. In this model, the process of transcription opens up the DNA double helix, enhancing the probability of mutation in the transcribed strand, but only mutations on the transcribed strand are transferred to mRNA and translate into mutants proteins that explore novel activities. Subsequently, the activity of the mutant protein may enable the cell to replicate and thereby fixes the initial adaptive mutation on both DNA strands. In agreement with the retromutagenesis model, *lacZ* amber mutations on the transcribed strand were recently isolated in approximately 10-fold excess over mutations on the non-transcribed strand upon treatment with the mutagen nitrous acid [3]. However, the prevalence of this molecular mechanism has not been studied in a non-mutator background and has not been validated in a whole genome context.

To gain new, in-depth knowledge of the mechanisms behind adaptive mutation we wanted to study the genetic changes in a background as undisturbed as possible. For this purpose, we designed an *E. coli* strain incapable of fermenting maltose, plated it on rich medium with maltose, and subsequently collected all mutants starting to outgrow on colonies over the course of two months (Figure 1).

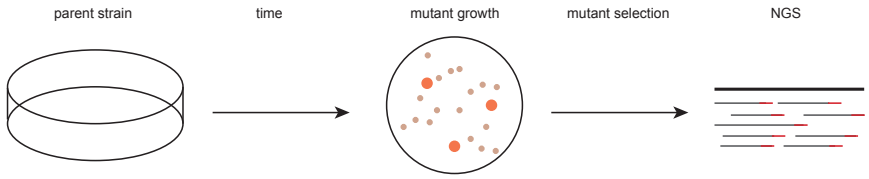


Figure 1. Schematic illustration of the experimental set-up. The parental strain is incapable of using the carbon source available in the plates, but after some days on plates, mutants adapted to their environment appear. These mutants are collected and subjected to next-generation genome sequencing (NGS).

Results

Aging colonies give rise to mutants in a non-mutator-based experimental system

Our starter strain is a derivative of the model bacterium *Escherichia coli* K12 MG1655 with its *cyaA* gene deleted, precluding synthesis of the signalling molecule cyclic AMP. As a consequence, a great many genes involved in carbon source catabolism cannot be expressed because they belong to operons requiring the cAMP Receptor Protein (CRP) complex to be activated [4]. These cells can grow for some time on rich media, but, after having used the accessible sources present in the available medium (MacConkey medium), the cells cannot grow further but remain as small colonies caught in a quiescent state.

Around four days after plating, mutants capable of fermenting maltose started to appear as red papillae-like structures on the white colonies, and these mutants continued to appear over the next two months as some cells adapted to their environment. Mutant papillae outgrew on approximately one in 200 colonies, and progressively invaded the surface of the plate (Figure 2a) making it necessary to start with a large number of plates to be able to sample mutants at the late time points. All mutant strains were purified and displayed a variety of phenotypes on different carbon sources (Supplementary Table 1).

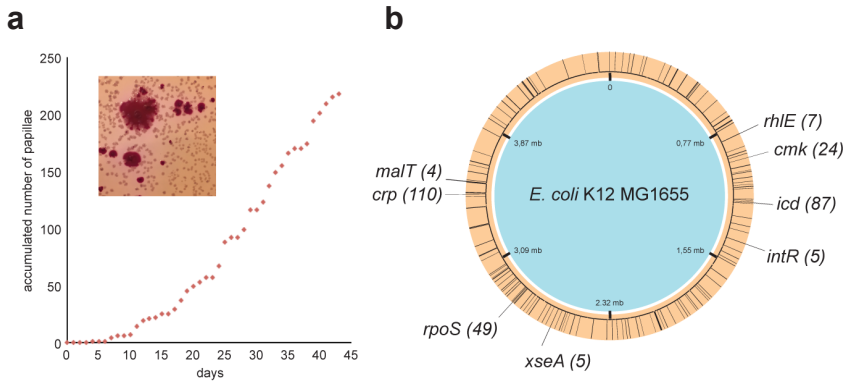


Figure 2. When subjected to prolonged incubation, *Escherichia coli* K12 MG1655 *cyaA*- cells adapt to their environment. (a) *E. coli cyaA*- cease to grow after forming small white colonies, when maltose is the major carbon source. After 4- 5 days of incubation, mutants start occurring as red papillae on MacConkey agar (inserted image) and continue to form over approximately 6 weeks. (b) 96 mutant papillae were isolated, phenotypically characterized and genome sequenced. Loci with less than three mutations are uniformly distributed on the circular genome as illustrated with black lines. Hotspots of genes with more than two mutations are indicated with double-sized black lines and the total number of mutations in each gene is stated within brackets.

Adaptive mutations are located in specific hotspot genes

We sequenced the genomes of 96 mutants, spanning the whole two-month period, and identified an average of four to five mutations per genome, the majority localized in a few hotspots only (Figure 2b and Supplementary Table 2). The hotspots are located in the following genes (Figure 2b): *cmk* (cytidylate kinase), *crp* (cyclic AMP receptor protein), *malT* (activator of the maltose regulon transcription), *rhlE* (ATP-dependent DNA helicase), *rpoS* (sigma factor sigma 38), *xseA* (large subunit of exonuclease VII), and two loci that are known to be unstable in laboratory strains; *e14-icd* (e14 prophage inserted in the NADP-dependent isocitrate dehydrogenase gene) and *intR* (Rac prophage integrase).

From the way the experiment was constructed, it could be expected that mutations in the *crp* gene would account for growth on maltose in a *ΔcyaA* background. However, the difficulty of obtaining cAMP-independent (*crp**) mutants witnessed by Jon Beckwith's laboratory suggested that under exponential growth such mutations were very rare (of the order of 1 in 10^9 cells [5]). In light of this, the ease with which we obtained such mutations in resting colonies was utterly unexpected.

It has been repeatedly observed that in *E. coli* laboratory strains, the *rpoS* gene is often inactivated [6,7]. This is also what we observed at the *rpoS* hotspot (point mutations in general, but also frameshifts and a deletion of the region, see Supplementary Table 2). Inactivation of the RpoS protein may have triggered an increase in age-related mutagenesis, as this transcription factor is involved in oxidative stress response during stationary phase [8]. However, before we

understand the molecular mechanism, we can only speculate why the remaining hotspot genes are targeted.

The majority of adaptive mutations take place on the transcribed strand

Theoretically, 12 different types of mutations are possible in DNA (A-C/G/T, C- A/G/T, G-A/C/T or T-A/C/G), but it is established that, due to respiration under stationary phase conditions, 8-oxo-G induced mutations are the most frequent to occur in the absence of an additional mutagenic process [9,10]. In line with this, a dominant proportion of the mutations that were observed during the generation of the carbon-positive papillae is consistent with the involvement of 8-oxo-G (G-T and C-A transversions, 69% of the total number of missense mutations, Figure 3a). A large proportion of the remaining mutations (28% of the total number of missense mutations identified) are likely the result of cytosine deamination (G-C to A-T transitions, Figure 3a), another common mutational event in non-dividing cells [9,11].

Remarkably, the mutated base is highly dominantly present on the transcribed strand within gene coding sequences (84% for G transversions and 93% for C deaminations, Figure 3a). Even more extremely, in a total of 594 *crp* variants we sequenced from different papillae, 99% of the G transversion and C deamination events had taken place on the transcribed strand (Figure 3b). These observations are consistent with an increased mutation rate in transcribed regions (see Supplementary Table 2) and strongly support the retromutagenesis model (Figure 3 c-d) [1,3,12] – here in the absence of a mutagen. Importantly, the extreme strand bias considerably reduces the relevance of potential residual replication and growth in

the aging colonies – a heavily debated, but hard to test, part of the controversy of adaptive mutations [13].

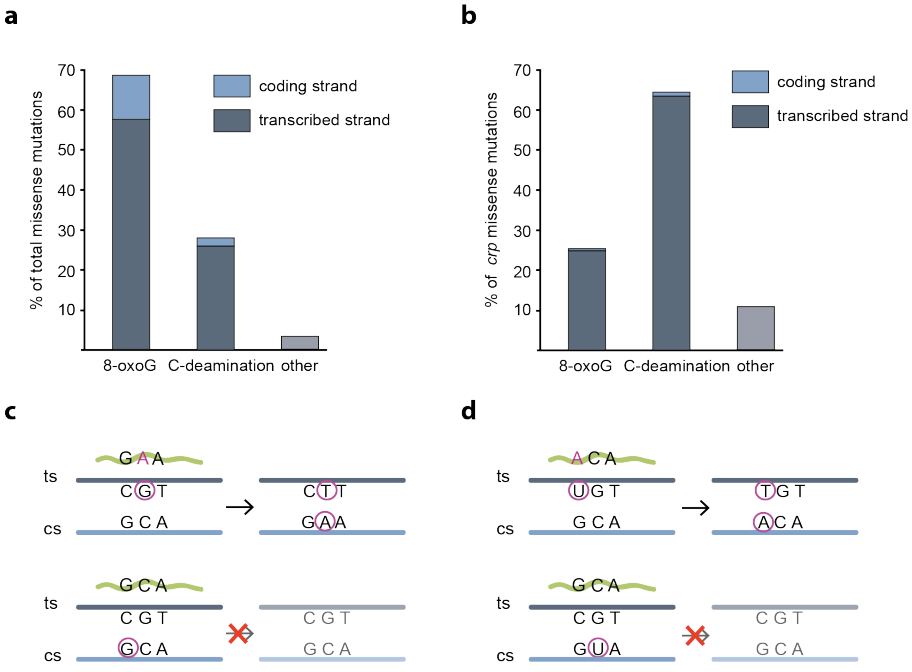


Figure 3. 8-oxoguanine formation and cytosine deamination are highly dominating on the transcribed strand in isolated papillae. (a) Assuming that all identified G-T and C-A transversions are the result of 8-oxoguanine formation, and all G-C to A-T transitions are the result of cytosine deamination, we analysed the localisation of the corresponding G- and C-residues in the 96 sequenced mutant genomes and in (b) the 594 sequenced *crp* sequences. Only missense mutations were analysed. (c) Upper panel: Illustration of the mechanism of retromutagenesis after 8-oxoguanine (encircled G) formation on the transcribed strand (ts, dark grey line) that enables insertion of an adenine ribonucleotide in the mRNA (green waved line). The mutated mRNA enables growth and a round of replication that permanently fixes the mutation on the coding strand (cs, light blue line) in daughter cells. Lower panel: 8- oxoguanine formation on the coding strand does not transfer to daughter cells. (d) Illustration of retromutagenesis after

cytosine nucleotide deamination into uracil (encircled U), enabling base pairing with A. Upper and lower panels show result of cytosine deamination on transcribed and coding strand respectively, as detailed in (c).

In the retromutagenesis model, mutations generated on the transcribed strand are selected for. In contrast, a so-called hitchhiking mutation is per definition not selected. Are the other hotspots selected for or hitchhikers? It is interesting to compare the specific nucleotide changes in *crp* that occur alone (singles) with those found only in combination (paired) with a *crp** mutation: 94% (399 out of 423) of single mutations were G transversion and C deamination events on the transcribed strand, whereas only 40% (27 out of 67) of the paired mutations were of this type. Furthermore, the 1% G transversions and C transitions found on the non-transcribed strand (eight events in total) were all in the paired positions. Thus, in contrast to the *crp** mutations, it is not possible to decipher the mechanism leading to these extra mutations in *crp*.

C deamination events outside *crp* are evenly distributed (50% on the transcribed and the non-transcribed strand). This indicates that the other C deamination events are not caused by retromutagenesis. Apart from events in *crp*, G transversion mutations were identified in the *xseA*, *cmk*, *malT* and *rpoS* hotspots, and 20 out of 21 *cmk* and 5 out of 5 *xseA* G transversions had taken place on the transcribed strand. This may indicate a selective advantage of these mutants, but hitchhiking is still theoretically possible because the *xseA* and *cmk* genes are transcribed from the same (+) strand as *crp* on the genome and thus may be fixed in the same round of replication as the *crp** mutants are. In contrast, *rpoS* mutations must have a selective advantage: 16 out of 17 *rpoS* G transversion missense mutations occur on the transcribed strand, but the *rpoS* gene is placed on the

complement (-) strand and thus retromutagenesis must have taken place independently of *crp** generation. Four out of four identified G transversion mutations in the *malT* hotspot were all placed on the non-transcribed strand and thus cannot have been selected by retromutagenesis.

Formation of adaptive mutations shows age-dependence

Time is an important factor in the development of mutant papillae and the sampling of adaptive mutants over two months enabled us to observe interesting age-related trends. Firstly, a minimum of four days of incubation is required for adaptive mutations to occur in our experimental setup. Secondly, the frequency of papillae with *rpoS* mutations increases over time, suggesting an increasingly selective advantage with age (Figure 4a). Alternatively, the mechanism(s) leading to *rpoS* mutations are becoming more prominent over time. Thirdly, G transversion mutations in *crp* increases over time, paralleled by a decrease in C-deaminations (Figure 4b).

Our observations are not consistent with a general increase in mutagenesis, as the frequency of rifampicin resistance is of the order of four mutants per 108 cells as generally observed [14]. Furthermore, removal of RecA, a previously suggested key actor in adaptive mutations [15], displayed only a marginally negative effect on the total number of papillae formed (Supplementary Figure 1).

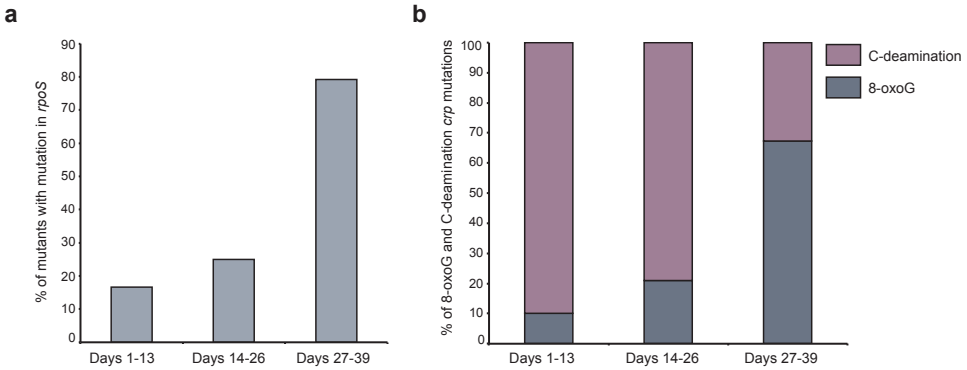


Figure 4. Time-dependent trends in mutational mechanisms. (a) The frequency of *rpoS* mutations in isolated mutants is increasing with time. The x-axis shows days after plating. (b) Isolated mutants show a distinct mutation pattern in the *crp* gene. Over time, there is a steady increase in 8-oxoG mutations whereas the C-deamination associated mutations decrease.

Discussion

In the present study, we sequenced 96 adaptive mutants isolated during a two-month period, after having arisen from a minimally perturbed genetic background. To our knowledge, a similar experiment has not been performed at this time scale previously. Importantly, time was an essential factor not only for the development of adaptive mutation, but also seemingly for the mutational mechanism and we show that a highly dominating proportion of the adaptive mutations are 8-oxo-G modifications on the transcribed strand. This strongly supports the retromutagenesis model as a major mechanism behind adaptive mutations. Were this process generalised to multicellular organisms it would be a fertile contribution to the initiation of cancer that parallels ageing.

Methods

Because the *fnr* gene is a homologue of *crp*, we used a starter strain previously identified as MG1655 at the ECGSC1 shown to harbour a

ca. 40 kb deletion in the *fnr* region [16]. The *cyaA* deletion was introduced with lambda Red-induced recombination [17]. In pilot experiments, we had noticed that the presence of a pTrc99A plasmid (Pharmacia Biotech) carrying the *E. coli tig* gene (PCR amplified with the oligonucleotides 5'CATGCCATGGTGAGGTAACAAGATGCAAGTTTC3' and 5'CGCGGATCCAATTACGCCTGCTGGTTCATC3' and cloned into the *Nco*I and *Bam*HI restriction sites) produced twice as many mutants that in absence of the *tig* plasmid, typically allowing us to recover around 200 mutants in each two months experiment, a figure quite convenient to get significant observations. Thus K12 Mg1655 (*cyaA*::*cat* delta *fnr*, pTrc*tig*) is our starter strain. The *recA* deletion from the Keio collections [18] was used to prepare a P1 lysate by standard procedures.

For growth on plates the lactose-free MacConkey medium was used [19] (Difco MacConkey Agar Base) supplemented with 1% carbohydrate (maltose in most experiments) or glycerol and chloramphenicol, 5 mg/litre, ampicillin, 100 mg/litre, and 1 mM IPTG. For subsequent 3-times purification of papillae, chloramphenicol, ampicillin and IPTG were omitted from the plates. To obtain isolated colonies on the plates, early stationary-phase grown bacteria from the LB liquid culture (37°C) were diluted in sterile water containing 9 g/l sodium chloride, to the concentration of 2.5×10^4 bacteria/ml and 100 μl of the bacterial suspension was spread onto MacConkey plates. The plates were subsequently placed in plastic boxes containing beakers with water to ensure constant humidity and placed for the duration of the experiment in an incubator at 37°C. To test for mutator phenotypes, rifampicin agar plates were prepared using the LB medium agar supplemented with 100 mg/liter rifampicin

prepared in methanol. At all times the plates were wrapped in aluminium foil for protection from light.

The genomic libraries were generated using the TruSeq®Nano DNA LT Sample Preparation Kit (Illumina Inc.). 498 independent papillae were selected for amplification of the *crp* region by direct colony PCR. The PCR reaction was made using Red-Taq polymerase (Sigma-Aldrich) according to the manufacturer's instructions (hybridisation at 55°C) with the following primers: forward primer 5′TTTCGGCAATCCAGAGACAGC3′ and reverse primer 5′AACATAGCACCAGCGTTTGTCG3′. The amplified *crp* regions were sequenced by Sanger method with two primers: forward 5′TTATCTGGCTCTGGAGAAAGC3′ and reverse primer 5′TCGAAGTGCATAGTTGATATCGG3′.

References

1. Bridges, B. A. mutY 'directs' mutation? *Nature* **375,** 741 (1995).
2. Doetsch, P. W. Translesion synthesis by RNA polymerases: occurrence and biological implications for transcriptional mutagenesis. *Mutat. Res.* **510,** 131– 140 (2002).
3. Morreall, J. *et al.* Evidence for retromutagenesis as a mechanism for adaptive mutation in Escherichia coli. *PLoS Genet* **11,** e1005477 (2015).
4. You, C. *et al.* Coordination of bacterial proteome with metabolism by cyclic AMP signalling. *Nature* **500,** 301–306 (2013).
5. Sabourin, D. & Beckwith, J. Deletion of the Escherichia coli crp gene. *J Bacteriol* **122,** 338–340 (1975).
6. Saxer, G. *et al.* Mutations in global regulators lead to metabolic selection during adaptation to complex environments. *PLoS Genet* **10,** e1004872 (2014).
7. Bleibtreu, A. *et al.* The rpoS gene is predominantly inactivated during laboratory storage and undergoes source-sink evolution in Escherichia coli species. *J Bacteriol* **196,** 4276–4284 (2014).
8. Nystrom, T. The free-radical hypothesis of aging goes prokaryotic. *Cell Mol Life Sci* **60,** 1333–1341 (2003).
9. Hall, B. G. Spectrum of mutations that occur under selective and non-selective conditions in E. coli. *Genetica* **84,** 73–76 (1991).
10. Bridges, B. A. Mutation in resting cells: the role of endogenous DNA damage. *Cancer Surv* **28,** 155–167 (1996).
11. Brégeon, D., Doddridge, Z. a., You, H. J., Weiss, B. & Doetsch, P. W. Transcriptional mutagenesis induced by uracil and 8-oxoguanine in Escherichia coli. *Mol. Cell* **12,** 959–970 (2003).
12. Davis, B. D. Transcriptional bias: a non-Lamarckian mechanism for substrate- induced mutations. *Proc Natl Acad Sci U S A* **86,** 5005–5009 (1989).
13. Maisnier-Patin, S. & Roth, J. R. The origin of mutants under selection: How natural selection mimics mutability (adaptive mutation). *Microb. Evol.* (2014).

14. Jin, D. J. & Gross, C. a. Mapping and sequencing of mutations in the Escherichia coli rpoB gene that lead to rifampicin resistance. *J. Mol. Biol.* **202,** 45–58 (1988).

15. Taddei, F., Halliday, J. A., Matic, I. & Radman, M. Genetic analysis of mutagenesis in aging Escherichia coli colonies. *Mol Gen Genet* **256,** 277–281 (1997).

16. Soupene, E. *et al.* Physiological studies of Escherichia coli strain MG1655: growth defects and apparent cross-regulation of gene expression. *J Bacteriol* **185,** 5611–5626 (2003).

17. Datta, S., Costantino, N. & Court, D. L. A set of recombineering plasmids for gram-negative bacteria. *Gene* **379,** 109–115 (2006).

18. Baba, T. *et al.* Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* **2,** 2006 0008 (2006).

19. MacConkey, A. Lactose-fermenting bacteria in feces. *J Hyg* **5,** 333–378 (1905).

Acknowledgements

This work was supported by The Novo Nordisk Foundation and a PhD grant from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme [FP7-People-2012-ITN], under grant agreement No. 317058, "BACTORY". We thank Ida Lauritsen and Emil Christian Fischer for technical assistance, and Anna Koza and Ida Bonde for their assistance with NGS.

Author contributions

AD and MHHN designed the experiments, supervised the work and wrote the manuscript. AZ and SW carried out the experiments, contributed to the manuscript and prepared the figures.

Competing financial interests statement

The authors have no competing financial interests.

Supplementary material to Paper I

Supplementary methods

Escherichia coli K12 remains the best-known living species. The genomes of several laboratory strains have been sequenced, and it has been observed that there is significant variation between strains [1]. Two major sources of laboratory strains have been investigated, MG1655, by Fred Blattner and co-workers [2], and W3110 by Hirotada Mori and co-workers (origin of the Keio collection [3]). Strain MG1655 seems to be the closest isolate from the original K12 isolate, which has now been lost. Beside a range of variants present in different laboratories, an initial isolate of MG1655 deposited at the *Escherichia coli* Genetic Stock Center (ECGSC) was shown to harbour a ca. 40 kb deletion in the fnr region [1]. For our experimental chassis we wished to construct a *cyaA* deletion strain, preventing cells from synthesising cyclic AMP. We expected that among adaptive mutants we might obtain *crp* derivatives of the *crp* gene (*crp**), coding for variants of the Cyclic AMP Receptor Protein (CRP), as previously obtained by Sabourin and Beckwith [4]. Because the FNR protein is strongly similar to CRP we were concerned that we might stumble on interference between *crp* and *fnr* that would have been confounding our observations. For this reason we decided to use as a chassis the strain previously identified as MG1655 at the ECGSC [1], a strain that was used with success recently in experiments exploring the swarming behaviour of *E. coli* [5]. In preliminary experiments we indeed found that papillae were produced on our selective medium (MacConkey plates supplemented with maltose as the superfluous carbon source). However, we wished to make the most of our experimental set-up, while we had noticed that many *E. coli* strains, when streaked on plates for conservation and then reused, had lost the activity of their *rpoS* gene [6–8]. RpoS has an important role under stationary phase conditions, in particular because it controls synthesis of a protein that should have an important role in our

experiment, the ribosome trigger factor (TIG), involved in the folding of proteins in *statu nascendi* [9]. We therefore introduced in our chassis, beside a replacement of the *cyaA* gene by a chloramphenicol cassette, a plasmid coding for *tig*. We noticed that the number of papillae produced under such conditions was about twice that in absence of the *tig* plasmid, allowing us to recover some 100 papillae or so in each two months experiment, a figure quite convenient to get significant observations. Our chassis is therefore strain AMBEC7001 (*cyaA*::*cat Δfnr*, pTrc*tig*).

In our set-up we expected that *crp** mutations would display a GASP phenotype. However, the experiments by Jon Beckwith's laboratory [4] showed that under exponential growth constitutively positive *crp* mutations were very rare (of the order of 1 in 10^9 cells). Mutagenesis would considerably increase the amount of such mutations. Hence, inactivating antimutator genes would result in an overall increase of mutations, without allowing much insight into the process. For this reason we were careful to avoid bacterial chasses that carry mutator genes. In particular, the MutM and MutY proteins correct mutations induced by the presence of 8-oxoguanine [10,11]; MutM removes 8-oxoG paired with C in DNA whereas MutY removes A paired with 8-oxoG in syn conformation in the double helix. As an essential requirement, the *mutM* and *mutY* genes are undamaged in our chassis. Indeed, our preliminary experiments showed that the strains submitted to ageing in stationary phase did not display an increase in rifampicin resistant mutants (*rpoB*).

The genomic libraries were generated using the TruSeq®Nano DNA LT Sample Preparation Kit (Illumina Inc.). Briefly, 100 ng of genomic DNA was diluted in 52.5 µl TE buffer and fragmented in Covaris Crimp Cap microtubes on a Covaris E220 ultrasonicator (Woburn).

According to Illumina's recommendations for a 350-bp average fragment size, the settings used were 5% duty factor, 175 W peak incident power, 200 cycles/burst, and 50-s duration under frequency sweeping mode at 5.5 to 6°C. The ends of fragmented DNA were repaired by T4 DNA polymerase, Klenow DNA polymerase, and T4 polynucleotide kinase. The Klenow exo minus enzyme was then used to tail an 'A' base to the 3' end of the DNA fragments. After ligation of adapters, DNA fragments ranging from 300 - 400 bp were recovered by bead purification. Finally, the adapter-modified DNA fragments were enriched by 3 cycles of PCR. The final concentration of each library was measured by Qubit® 2.0 Fluorometer and Qubit DNA Broad range assay (Life Technologies). Average dsDNA library sizes were determined using the Agilent DNA 7500 kit on an Agilent 2100 Bioanalyzer. Libraries were normalised and pooled in 10 mM Tris-HCl, pH 8.0, 0.05% Tween 20 to a final concentration of 10 nM. 10 pm pools of 20 libraries in 600 μl ice-cold HT1 buffer were denatured in 0.2N NaOH, loaded onto the flow cell provided in the MiSeq Reagent kit v2 and sequenced on a MiSeq (Illumina Inc.) platform, with a paired-end protocol and read lengths of 151 nt. The Illumina sequencing data was quality-trimmed using the Trimmomatic tool (version 0.32) [12] with the settings CROP:145 HEADCROP:15 SLIDINGWINDOW:4:15 MINLEN:30. The cleaned data was used as input for variant calling using the breseq pipeline (version 0.26.0) [13] with –j 4 and –b 20 as the only changes to default settings. The reference strain for this analysis was *Escherichia coli* MG1655 with the accession number NC_000913.3.

Supplementary discussion

Besides genes relevant to the process of carbon source usage, the sequencing of 96 papillae genomes revealed several hotspots and several regions of instability (see main text). The latter are all related

to mobile elements. In particular the *icd* gene is the place of prophage e14 attachment site [14], and it appears from our data that this region remains unstable in our experimental conditions. A similar observation was made when investigators explored the GASP phenotype [15]. In fact, the whole region is unstable as witnessed by mutations in genes *ymfE*, *ymfI* and *ymfJ* (see Supplementary Table 2). In addition to mutations in *crp*, *rpoS* and *malT*, gene *cmk*, coding for Cytidylate kinase, seems to be a mutational hotspot. In the *cmk rpsA* operon [16], the latter gene coding for ribosomal protein S1 in Proteobacteria and an mRNA binding protein in Firmicutes, is widely conserved in Bacteria [17]. Cytidylate kinase is critical for the production of CDP, the immediate precursor of dCDP that is an absolute requirement for the presence of cytosine in DNA [18], and a major source of thymine as well [19]. The metabolism during stationary phase may require maintenance of a stable CTP pool, asking for the presence of both CMK and nucleoside diphosphokinase, but the reason for an increased mutation number in the *cmk* gene is unknown. However, an unexpected phenotype of the *cmk* gene revealed that it was involved in the control of chromosome maintenance (*cmk* was originally named *mssA*, suppressor of a cold sensitive growth of mutants in the uridylate kinase gene [20]), suggesting that it may belong to some important role in this domain.

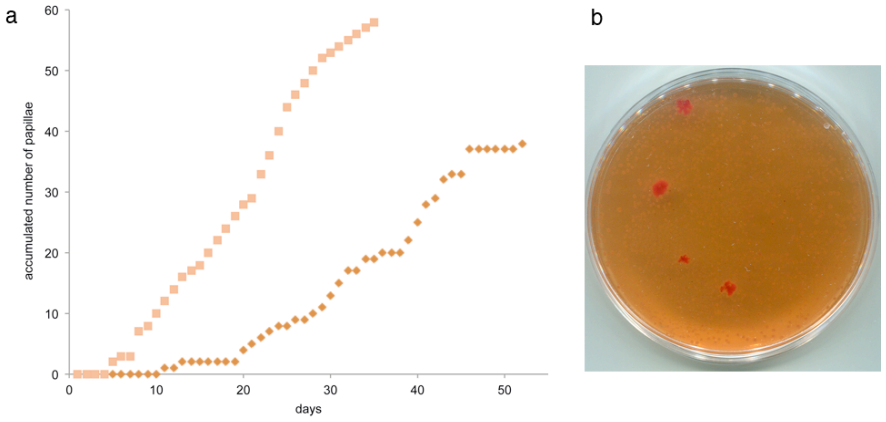
The present work resulted in the selection of a very large number of CRP variants. It seems that despite the number of mutants explored (ca. 600), there is still room for further mutants to appear. The phenotype of the papillae was assayed on MacConkey medium supplemented with maltose, mannitol, glycerol or sorbitol, as well as on EMB maltose plates (Supplementary Table 1). The corresponding phenotypes are driven by the sequence of the CRP protein, but

mutations elsewhere in the genome can also impact the phenotype (see for example in Supplementary Table 1 papilla 1 compared to papilla 3). The present work could be a first step in the understanding of functional networks directing carbon source utilisation. Remarkably, previous knowledge was far from covering the whole set of what we observed. This highlights a rarely underlined process of evolution that can be explored using experiments similar to the one presented in this article.

The two maltose-positive clones that carried a wild-type *crp* had a weak positive phenotype and a mutation in *rpoS*, besides other mutations. These mutants, in contrast to the mutants in the *crp* gene, did not grow on minimal plates supplemented with maltose. They may have uncovered pathways allowing for growth on components of the rich MacConkey medium that the starter strain could not use.

Finally, noting that we did not find any hint of RecA involvement in the genomes we sequenced, and to validate a possible role of RecA in adaptive mutation as previously suggested [21], we introduced a *recA* mutation in our root strain and monitored the appearance of positive papillae. Interestingly, while, overall, there does not seem to be a large decrease of the total number of papillae in a *recA*-background, the time course of their appearance is considerably retarded as compared to the situation observed in the *cyaA*- background (Supplementary Figure 1). Indeed, it takes a long time for the first papillae to appear in the *recA*- background, but their number keeps increasing, for up to two months, with the total number observed similar to that found in the *cyaA*- reference strain. In our context, *recA* mutants are growing more slowly than their parent counterpart. Whether this growth rate can account for the delay in papillae appearance warrants further studies.

Supplementary figure



Supplementary Figure 1. Time course of papillae formation in a *cyaA*-strain and *cyaA recA* double mutant. **(a)** Papillae formation is delayed in the *cyaA recA* (orange diamonds) mutant compared to the *cyaA* deficient reference strain (light orange squares). In the *cyaA* mutant the first papillae are appearing on the 5th day of experiment and terminating over the 32nd day, while in the *cyaA recA* double mutant the first papillae are appearing on the 10th day and keep appearing on the 52nd day after the inoculation. Papillae are represented as accumulated total number of occurrences. **(b)** An example of a typical plate with the papillae after 6 days.

References to Supplementary methods and discussion

1. Soupene, E. *et al*. Physiological studies of *Escherichia coli* strain MG1655: growth defects and apparent cross-regulation of gene expression. J Bacteriol 185, 5611–5626 (2003).
2. Blattner, F. R. *et al*. The complete genome sequence of *Escherichia coli* K-12. Science (80-.). 277, 1453–1462 (1997).
3. Hayashi, K. *et al*. Highly accurate genome sequences of *Escherichia coli* K-12 strains MG1655 and W3110. Mol Syst Biol 2, 2006 0007 (2006).
4. Sabourin, D. & Beckwith, J. Deletion of the *Escherichia coli crp* gene. J Bacteriol 122, 338– 340 (1975).
5. Liu, C. *et al*. Sequential establishment of stripe patterns in an expanding cell population. Science (80-.). 334, 238–241 (2011).
6. Dong, T., Chiang, S. M., Joyce, C., Yu, R. & Schellhorn, H. E. Polymorphism and selection of *rpoS* in pathogenic *Escherichia coli*. BMC Microbiol 9, 118 (2009).
7. Jishage, M. & Ishihama, A. Variation in RNA polymerase sigma subunit composition within different stocks of *Escherichia coli* W3110. J Bacteriol 179, 959–963 (1997).
8. Visick, J. E., Cai, H. & Clarke, S. The L-isoaspartyl protein repair methyltransferase enhances survival of aging *Escherichia coli* subjected to secondary environmental stresses. J Bacteriol 180, 2623–2629 (1998).
9. Silva, I. J., Ortega, A. D., Viegas, S. C., Garcia-Del Portillo, F. & Arraiano, C. M. An RpoS- dependent sRNA regulates the expression of a chaperone involved in protein folding. RNA 19, 1253–1265 (2013).
10. Michaels, M. L., Cruz, C., Grollman, A. P. & Miller, J. H. Evidence that MutY and MutM combine to prevent mutations by an oxidatively damaged form of guanine in DNA. Proc Natl 11. Acad Sci U S A 89, 7022–7025 (1992).
11. Tajiri, T., Maki, H. & Sekiguchi, M. Functional cooperation of MutT, MutM and MutY proteins in preventing mutations caused by spontaneous oxidation of guanine nucleotide in *Escherichia coli*. Mutat Res 336, 257–267 (1995).

12. 12. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120 (2014).
13. Deatherage, D. E. & Barrick, J. E. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. Methods Mol Biol 1151, 165–188 (2014).
14. Hill, C. W., Gray, J. A. & Brody, H. Use of the isocitrate dehydrogenase structural gene for attachment of e14 in *Escherichia coli* K-12. J Bacteriol 171, 4083–4084 (1989).
15. Zhang, Q. *et al*. You cannot tell a book by looking at the cover: Cryptic complexity in bacterial evolution. Biomicrofluidics 8, 52004 (2014).
16. Fricke, J., Neuhard, J., Kelln, R. A. & Pedersen, S. The *cmk* gene encoding cytidine monophosphate kinase is located in the *rpsA* operon and is required for normal replication rate in *Escherichia coli*. J Bacteriol 177, 517–523 (1995).
17. Nitschke, P. *et al*. Indigo: a World-Wide-Web review of genomes and gene functions. FEMS Microbiol Rev 22, 207–227 (1998).
18. Danchin, A. Comparison between the *Escherichia coli* and *Bacillus subtilis* genomes suggests that a major function of polynucleotide phosphorylase is to synthesize CDP. DNA Res 4, 9–18 (1997).
19. Weiss, B. & Wang, L. De novo synthesis of thymidylate via deoxycytidine in *dcd* (dCTP deaminase) mutants of *Escherichia coli*. J Bacteriol 176, 2194–2199 (1994).
20. Yamanaka, K., Ogura, T., Koonin, E. V, Niki, H. & Hiraga, S. Multicopy suppressors, *mssA* and *mssB*, of an *smbA* mutation of *Escherichia coli*. Mol Gen Genet 243, 9–16 (1994).
21. Taddei, F., Halliday, J. A., Matic, I. & Radman, M. Genetic analysis of mutagenesis in aging *Escherichia coli* colonies. Mol Gen Genet 256, 277–281 (1997).

Supplementary tables

Supplementary table 1.

Phenotypic characterisation of the 96 isolated mutants. The phenotypes of the papillae were assayed on MacConkey medium supplemented with maltose, mannitol, glycerol or sorbitol, as well as on EMB maltose plates. On MacConkey plates, colonies can be either red, pink or white, and in some cases red at the centre but white around ("fisheye" phenotype). On EMB, colonies can be very dark (black), often with a remarkable green shine (gold green), purple, or in some cases white.

Pap #	maltose	mannitol	glycerol	sorbitol	EMB	L40M	E56K	E56D	AAA->174	S63Y	T141R	T141K	A145T	A145E	Q171K	Q175K	M190K	M190I	Q194K
1	red	red	red	red	gold green	0	0	0	0	0	0	0	1	0	0	0	0	0	0
2	red	red	red	red	purple	0	0	0	0	0	1	0	0	0	0	0	0	0	0
3	white	red	red	white	white	0	0	0	0	0	0	0	1	0	0	0	0	0	0
4	red	red	red	red	purple	0	0	0	0	0	1	0	0	0	0	0	0	0	0
5	white	red	red	white	white	0	0	0	0	0	0	0	1	0	0	0	0	0	0
8	red	red	red	red	black	0	0	0	0	0	0	0	1	0	0	0	0	0	0
10	fisheye	red	red	red	gold green	0	0	0	0	0	0	0	1	0	0	0	0	0	0
14	white	red	red	red	gold green	0	0	0	0	0	0	0	1	0	0	0	0	0	0
15	red	red	red	red	black	0	0	0	0	0	0	0	1	0	0	0	0	0	0
16	red	red	red	red	gold green	0	0	0	0	0	0	0	1	0	0	0	0	0	0
18	red	red	red	red	purple	0	0	0	0	0	0	0	0	1	0	0	0	0	0
19	white	red	red	white	white	0	0	0	0	0	0	0	1	0	0	0	0	0	0
22	red	red	red	red	purple	0	0	0	0	0	0	0	0	1	0	0	0	0	0
23	red	red	red	red	gold green	0	0	0	0	0	0	0	1	0	0	0	0	0	0
24	red	red	red	red	black	0	0	0	0	0	0	0	1	0	0	0	0	0	0
27	red	red	red	red	black	0	0	0	0	0	0	0	1	0	0	0	0	0	0
28	red	red	red	red	gold green	0	0	0	0	0	0	0	1	0	0	0	0	0	0
31	fisheye	red	red	fisheye	purple	0	0	0	0	0	0	0	1	0	0	0	0	0	0
34	white	red	red	red	gold green	0	0	0	0	0	0	0	1	0	0	0	0	0	0
35	fisheye	red	red	red	gold green	0	0	0	0	0	0	0	1	0	0	0	0	0	0
36	white	red	red	fisheye	white	0	0	0	0	0	0	0	1	0	0	0	0	0	0
37	red	red	red	fisheye	black	0	0	0	0	0	0	0	1	0	0	0	0	0	0
39	red	red	red	fisheye	purple	0	0	0	0	0	0	0	1	0	0	0	0	0	0
42	white	red	red	red	white	0	0	0	0	0	0	0	1	0	0	0	0	0	0
43	red	red	red	fisheye	black	0	0	0	0	0	0	0	1	0	0	0	0	0	0
44	red	red	red	red	gold green	0	0	0	0	0	0	1	1	0	0	0	0	0	0
45	red	red	red	fisheye	gold green	0	0	0	0	0	0	0	1	0	1	0	0	0	0
48	red	red	red	white	gold green	0	0	0	0	0	0	0	1	0	0	0	0	0	0
49	red	red	red	red	purple	0	0	0	0	0	0	0	1	0	0	0	0	0	0
50	red	red	red	red	gold green	0	0	0	0	0	0	0	1	0	0	0	0	0	1
51	white	red	red	red	white	0	0	0	0	0	0	0	1	0	0	0	0	0	0
52	red	red	red	red	gold green	0	0	0	0	0	0	0	1	0	0	0	1	0	0

53	white	red	red	red	gold green	0	0	0	0	0	0	0	1	0	0	0	0	0	0
54	white	red	red	white	white	0	0	0	0	0	0	0	1	0	0	0	0	0	0
56	fisheye	red	red	fisheye	black	0	0	0	0	0	0	0	1	0	0	0	0	0	0
57	red	red	red	fisheye	black	0	0	1	0	0	0	0	1	0	0	0	0	0	0
58	white	red	red	red	black	0	0	0	0	0	0	0	1	0	0	0	0	0	0
59	red	red	red	red	gold green	0	0	0	0	0	0	0	0	1	1	0	0	0	0
60	red	red	red	fisheye	black	0	0	0	0	0	0	0	1	0	0	0	0	0	0
63	red	red	red	red	black	0	0	0	0	0	0	0	1	0	0	0	0	0	0
64	fisheye	red	red	red	gold green	0	0	0	1	0	0	0	0	0	0	0	0	0	0
65	red	red	red	fisheye	purple	0	0	0	0	0	1	0	0	0	0	0	0	0	0
67	white	red	red	red	gold green	0	0	0	0	0	0	0	1	0	0	0	0	0	0
69	fisheye	red	red	fisheye	gold green	0	0	0	0	0	0	0	0	1	0	0	0	0	0
82	fisheye	red	red	red	purple	0	0	0	0	0	0	0	0	1	0	0	0	0	0
87	white	red	red	red	white	0	0	0	0	0	0	0	1	0	0	0	0	0	0
88	fisheye	red	red	fisheye	black	0	0	0	0	0	0	0	1	0	0	0	0	0	0
90	fisheye	red	white	fisheye	black	0	0	0	0	0	0	0	1	0	0	0	0	0	0
93	red	red	red	fisheye	gold green	0	0	0	0	0	0	0	1	0	0	0	0	0	0
95	white	red	red	white	white	0	0	0	0	0	0	0	1	0	0	0	0	0	0
96	red	red	red	fisheye	black	0	0	0	0	0	0	0	1	0	0	0	0	0	0
97	red	red	red	red	purple	0	0	0	0	0	0	0	1	0	0	0	0	0	0
98	red	red	red	red	black	0	0	0	0	0	0	0	1	0	0	0	0	0	0
101	red	red	white	red	purple	0	0	0	0	0	0	0	0	1	0	0	0	0	0
102	red	red	red	white	black	0	0	0	0	0	0	0	1	0	0	0	0	1	0
105	white	red	white	fisheye	white	0	0	0	0	0	0	0	0	1	0	0	0	0	0
106	red	red	red	fisheye	black	0	0	0	0	0	0	0	1	0	0	0	0	0	0
112	white	red	red	fisheye	purple	0	0	0	0	0	0	0	1	0	0	0	0	0	0
117	white	pink	white	red	white	0	0	0	0	0	0	0	0	0	0	0	0	0	0
118	white	red	red	fisheye	white	0	0	0	0	0	0	0	0	1	0	0	0	0	0
119	red	red	red	red	gold green	0	1	0	0	0	0	0	0	1	0	0	0	0	0
120	white	red	red	red	white	0	0	0	0	0	0	0	1	0	0	0	0	0	0
122	red	white	white	white	white	0	0	0	0	0	0	0	0	0	0	0	0	0	0
124	white	red	red	red	purple	0	0	0	0	0	0	0	0	1	0	0	0	0	0
125	white	red	white	red	white	0	0	0	0	0	0	0	0	1	0	0	0	0	0
126	red	red	white	red	black	0	0	0	0	0	0	0	0	1	0	0	0	0	0

128	red	red	red	white	purple	0	0	0	0	0	0	0	1	0	0	0	0	0	0
129	red	red	red	fisheye	black	0	0	0	0	0	0	0	1	0	0	0	0	0	0
132	white	red	white	fisheye	purple	0	0	0	0	0	0	0	0	1	0	0	0	0	0
138	red	red	red	fisheye	purple	0	0	0	0	0	0	0	0	1	0	0	0	0	0
139	white	red	red	fisheye	purple	0	0	0	0	0	0	0	0	1	0	1	0	0	0
141	red	red	red	red	black	0	0	0	0	0	0	0	0	1	0	0	0	0	0
146	red	red	red	fisheye	purple	0	0	0	0	0	1	0	0	0	0	0	0	0	0
147	white	red	red	fisheye	purple	0	0	0	0	0	0	0	0	1	0	0	0	0	0
149	white	red	red	red	white	0	0	0	0	0	0	0	0	1	0	0	0	0	0
150	red	red	red	red	purple	0	0	0	0	0	0	0	0	1	0	0	0	0	0
153	red	red	red	fisheye	purple	0	0	0	0	0	0	0	0	1	0	0	0	0	0
154	white	red	red	fisheye		0	0	0	0	0	0	0	1	0	0	0	0	0	0
157	red	red	red	red	black	0	0	0	0	0	0	0	0	1	1	0	0	0	0
163	red	red	red	red	black	0	0	0	0	0	0	0	0	1	0	0	0	0	0
164	white	red	red	red	white	0	0	0	0	0	0	0	1	0	0	0	0	0	0
166	white	red	red	red	purple	0	0	0	0	0	0	1	0	1	0	0	0	0	0
168	red	red	red	fisheye	purple	0	0	0	0	0	0	0	0	1	0	0	0	0	0
169	white	red	red	fisheye	white	0	0	0	0	0	0	0	1	0	0	0	0	0	0
172	red	red	red	fisheye	purple	0	0	0	0	0	0	0	0	1	0	0	0	0	0
175	red	red	red	fisheye	black	1	0	0	0	0	0	0	0	1	0	0	0	0	0
176	fisheye	red	red	white	purple	0	0	0	0	0	0	0	0	1	0	0	0	0	0
177	red	red	red	white	purple	0	0	0	0	0	0	0	0	1	0	0	0	0	0
179	white	red	red	red	white	0	0	0	0	0	0	1	0	1	0	0	0	0	0
180	red	red	red	red	gold green	0	0	0	0	1	0	0	0	1	0	0	0	0	0
182	fisheye	red	red	red	white	0	0	0	0	0	0	0	1	0	0	0	0	0	0
184	red	red	red	white	black	0	0	0	0	0	0	0	0	1	0	0	0	0	0
185	white	red	red	fisheye	white	0	0	0	0	0	0	0	1	0	0	0	0	0	0
186	white	red	red	white	purple	0	0	0	0	0	0	0	0	1	0	1	0	0	0
188	white	red	red	red	purple	0	0	0	0	0	0	0	0	1	0	0	0	0	0
190	red	red	red	red	black	0	0	0	0	0	0	0	0	1	1	0	0	0	0

Supplementary table 2.

Mutations in the 96 genomes sequenced using NGS. The mutation sites observed in the 96 mutants are described. The origin of the mutation is colour coded: G->T transversions are in blue while C->A transversions are in green; C->T transitions are in yellow and G->A transitions are in pink. Hotspot regions are coloured light grey (and the *crp* region is in orange).

type	location	gene name	change	annotation	total count	Days 1-13	Days 14-26	Days 27-39
SNP	30652	carA	C->A	D334E	1	0	0	1
SNP	70573	araC	G->T	187	1	0	1	0
SNP	106044	ftsZ	C->A	P247Q	1	0	0	1
SNP	127795	aceF/lpd	C->A	intergenic (+208/-117)	1	0	0	1
SNP	177157	yadS	G->T	468	1	0	0	1
SNP	194684	dxr	C->A	1164	1	0	0	1
MCDEL	224944	rrsH	3994 bp		1	1	0	0
SNP	258698	crl	C->A	pseudogene (23/331 nt)	1	0	1	0
SNP	292555	yagK	G->T	L132I	1	0	0	1
SNP	314497	eaeH	G->T	pseudogene (141/872 nt)	1	0	0	1
SNP	326495	betA	G->T	N251K	1	0	0	1
SUB	366351	lacZ/lacI	2 bp ->TT		1	0	0	1
SNP	366409	lacZ/lacI	C->T	intergenic (-104/+19)	1	0	0	1
SNP	367573	lacI/mhpR	G->A	intergenic (-63/+14)	1	0	0	1
SNP	420365	brnQ	C->A	Q259K	1	0	0	1
SNP	487471	mscK	G->T	M312I	1	0	0	1
DEL	494093	ybaB	1 bp		1	0	0	1
INS	503380	ybaL/fsr	->A	intergenic (-142/+96)	1	0	1	0
SNP	557572	folD	G->T	A57D	1	0	1	0
SNP	624066	fepB	C->A	D149Y	1	0	1	0
SNP	699604	nagD	G->T	723	1	0	1	0
SNP	726568	kdpB	C->A	A157S	1	0	0	1
DEL	728596	kdpA	1 bp		1	0	0	1
SNP	751158	ybgQ	G->T	R338S	1	0	0	1
SNP	754644	gltA/sdhC	G->T	intergenic (-176/-533)	1	0	0	1
SNP	766215	mngA	G->T	A78S	1	0	0	1
SNP	830395	ybiH	G->T	249	1	0	1	0
MOB	831184	rhlE	IS1 1 9 bp		7	1	2	4

SNP	900818	artJ/artM	G->T	intergenic (-243/+48)	1	0	1	0
SNP	923622	clpA	C->A	A120E	1	0	0	1
SNP	938191	rarA	C->A	198	1	0	0	1
SNP	961307	cmk	C->A	107	2	1	0	1
SNP	961511	cmk	C->A	A104E	1	0	0	1
SNP	961669	cmk	C->A	R157S	1	0	1	0
SNP	961687	cmk	C->T	487	1	0	0	1
SNP	961746	cmk	C->A	D182E	1	0	1	0
SNP	961847	cmk	C->A	A216E	17	4	7	6
SNP	961857	cmk	C->A	657	1	0	0	1
DEL	1001023	elfC	1 bp		1	0	0	1
SNP	1026791	mgsA	G->T	225	1	0	1	0
SNP	1034672	hyaB	G->T	1419	1	0	1	0
SNP	1062381	torD	G->T	D195Y	1	1	0	0
SNP	1063291	cbpA	A->G	V162A	1	0	0	1
SNP	1085045	phoH	C->A	54	1	0	1	0
SNP	1100367	ycdZ	C->A	72	1	1	0	0
SNP	1112357	opgH	C->A	R499S	1	0	0	1
SNP	1121659	pyrC	G->T	R317S	1	0	0	1
SNP	1179118	nagK	C->A	R176S	1	0	0	1
SNP	1184194	potB	G->T	P84Q	1	1	0	0
SNP	1192280	hflD	G->T	R118S	1	0	0	1
SNP	1196220	icd	C->T	1098	19	1	9	9
SNP	1196232	icd	C->T	1110	19	1	9	9
MCDEL	1196232	icd	15224 bp		1	0	0	1
MCDEL	1196243	icd	15224 bp		1	0	1	0
SNP	1196245	icd	T->C	1123	17	1	8	8
SNP	1196247	icd	A->G	1125	17	1	8	8
MCDEL	1196254	icd	15177 bp		1	0	0	1

MCDEL	1196266	icd	15196 bp		1	0	1	0
MCDEL	1196269	icd	15172 bp		1	0	0	1
MCDEL	1196269	icd	15187 bp		1	0	0	1
MCDEL	1196270	icd	15162 bp		1	0	1	0
MCDEL	1196273	icd	15164 bp		1	0	0	1
MCDEL	1196273	icd	15166 bp		1	1	0	0
MCDEL	1196274	icd	15154 bp		1	0	0	1
SNP	1196277	icd	C->T	1155	6	0	4	2
MCDEL	1196277	icd	15174 bp		1	0	0	1
MCDEL	1196277	icd	15182 bp		1	0	1	0
MCDEL	1196277	icd	15180 bp		1	0	1	0
SNP	1196280	icd	G->C	1158	5	0	3	2
MCDEL	1196280	icd	15172 bp		1	0	1	0
SNP	1196283	icd	A->G	1161	4	0	3	1
MCDEL	1196283	icd	15150 bp		1	0	0	1
MCDEL	1196286	icd	15159 bp		1	0	1	0
MCDEL	1196287	icd	15158 bp		1	0	1	0
MCDEL	1196289	icd	15160 bp		1	0	0	1
SNP	1196292	icd	C->T	1170	1	0	1	0
MCDEL	1196297	icd	15127 bp		1	0	1	0
MOB	1197566	ymfE	IS1 1 9 bp		1	0	0	1
SNP	1201771	ymfI	C->A	A92D	1	0	1	0
SNP	1202203	ymfJ/cohE	C->A	intergenic (-119/+56)	1	0	1	0
SNP	1278166	narK	G->T	210	1	0	1	0
DEL	1299499	ychE/oppA	1199 bp		15	1	6	8
MCDEL	1299499		1198 bp		15	1	6	8
SNP	1322080	trpE	G->T	867	1	0	1	0
SNP	1355233	sapB	G->T	L80M	1	0	0	1
MOB	1413068	intR	IS30 -1 2 bp		5	0	2	3

SNP	1413735	rcbA	A->T	6	1	0	1	0
SNP	1461476	paaJ	C->A	A195E	1	0	1	0
MOB	1496976	opgD	IS2 1 5 bp		1	0	0	1
SNP	1523535	yncE	G->A	V77I	1	0	1	0
SNP	1538739	narY	C->A	A39S	1	1	0	0
SNP	1617812	ydeA	G->T	S262I	1	0	0	1
SNP	1619107	marC/marR	C->T	intergenic (-199/-13)	1	0	0	1
MCDEL	1632872	ydfJ/ydfK/pinQ/tfaQ/stfQ	2085 bp		1	0	0	1
SNP	1667788	mlc	G->T	777	1	0	0	1
SNP	1738585	purR	G->T	D248Y	1	0	0	1
INS	1751128	ydhW	->ATAAGCC	coding (597/648 nt)	1	0	0	1
SNP	1761901	sufC	G->T	612	1	0	1	0
SNP	1801238	thrS	G->T	L445M	1	0	0	1
SNP	1831528	astC	G->T	P152Q	1	0	1	0
SNP	1832172	astC/xthA	G->T	intergenic (-190/-256)	1	0	1	0
SNP	1841349	ynjF	G->T	55	1	0	1	0
MOB	1879829	yeaR	IS186 -1 6 bp		1	0	1	0
SNP	1881758	dmlR	G->T	R18S	1	0	0	1
SNP	1895250	pabB	G->T	R149L	1	0	1	0
MCDEL	1915128	proQ	4 bp		1	0	0	1
DEL	1915128	proQ	5 bp		1	0	0	1
SNP	1915297	proQ	G->T	R80S	1	0	1	0
SNP	1967724	cheB	C->A	V260L	1	0	1	0
SNP	2014743	fliF	C->T	1515	1	0	0	1
SNP	2030770	vsr	G->T	P50Q	1	0	0	1
SNP	2046102	yeeJ	G->A	G389S	1	0	1	0
MCDEL	2066261	insH1	953 bp		1	0	1	0
SNP	2079224	yeeX	G->T	D46E	1	0	1	0
SNP	2142922	udk	G->T	27	1	0	1	0

SNP	2165322	yegP	C->T	S45L	1	0	1	0
SNP	2166291	yegQ	A->G	624	1	0	0	1
SNP	2199955	yehH	C->A	pseudogene (634/948 nt)	1	0	0	1
SNP	2214426	yehU	C->A	219	1	1	0	0
AMP	2229165	yohP/dusC			2	0	1	1
MCDEL	2230197	dusC	7 bp		1	0	1	0
DEL	2230197	dusC	8 bp		1	0	1	0
SNP	2249250	yeiE	G->T	S123R	1	0	0	1
INS	2278761	bcr	->C	coding (1000/1191 nt)	1	0	0	1
SNP	2305176	mqo	C->A	D527Y	1	0	1	0
SNP	2336249	yfaA	C->A	396	1	0	1	0
SNP	2415034	pta	C->A	H96Q	1	0	0	1
SNP	2495925	lpxP	G->T	G94V	1	0	1	0
SNP	2528679	ligA	G->T	L500M	1	0	1	0
SNP	2561598	yffM	C->A	231	1	0	1	0
INS	2575843	eutS/maeB	->CA	intergenic (-38/+255)	1	0	0	1
SNP	2586389	narQ	G->T	S220I	1	0	0	1
SNP	2611203	hyfI	G->T	498	1	0	0	1
SNP	2635597	xseA	C->A	H456N	5	0	2	3
SNP	2661961	iscR	G->T	S57R	1	0	0	1
SNP	2682932	yphH	C->A	70	1	0	1	0
SNP	2708553	rseB	G->A	R68C	1	0	0	1
SNP	2762396	yfjK	G->T	A382D	1	0	0	1
SNP	2777847	ypjF/ypjA	G->T	intergenic (+65/+299)	1	0	1	0
SNP	2796042	csiR	G->T	M123I	1	0	1	0
SNP	2816171	gshA	G->T	T90K	1	0	0	1
DEL	2827989	[srlD][nlpD]	40677 bp		1	0	0	1
MCDEL	2827989	[srlD][nlpD]	40676 bp		1	0	0	1
SNP	2854843	fhlA	G->T	S169I	1	0	0	1

SNP	2860223	pphB	C->A	A155E	1	0	0	1
INS	2866592	rpoS	->C	coding (960/993 nt)	1	0	0	1
SNP	2866630	rpoS	C->A	922	1	0	0	1
MCDEL	2866646	rpoS	5 bp		1	0	0	1
DEL	2866646	rpoS	6 bp		1	0	0	1
SNP	2866654	rpoS	C->A	898	1	0	1	0
SNP	2866703	rpoS	G->T	849	3	0	1	2
SNP	2866783	rpoS	G->A	769	1	0	0	1
DEL	2866806	rpoS	1 bp		1	1	0	0
DEL	2866812	rpoS	1 bp		1	0	0	1
SNP	2866825	rpoS	C->A	727	1	1	0	0
SNP	2866890	rpoS	G->T	662	2	0	2	0
SNP	2866978	rpoS	C->A	574	1	0	0	1
DEL	2866982	rpoS	1 bp		1	0	0	1
INS	2866999	rpoS	->A	coding (553/993 nt)	1	0	0	1
SNP	2867018	rpoS	G->T	534	1	0	0	1
MOB	2867036	rpoS	IS1 -1 9 bp		1	0	1	0
SNP	2867050	rpoS	G->T	P168T	1	0	0	1
DEL	2867056	rpoS	1 bp		1	0	0	1
INS	2867121	rpoS	->AATCCACCAGGTTGCGTATG	coding (431/993 nt)	1	0	0	1
SNP	2867163	rpoS	G->T	A130E	1	0	0	1
INS	2867164	rpoS	->TCTACCGC	coding (388/993 nt)	1	0	0	1
DEL	2867170	rpoS	1 bp		1	0	1	0
SNP	2867176	rpoS	C->A	G126W	1	0	0	1
SNP	2867182	rpoS	T->A	N124Y	1	0	0	1
MCDEL	2867245	rpoS	11 bp		1	0	0	1
DEL	2867245	rpoS	12 bp		1	0	0	1
SNP	2867258	rpoS	G->T	N98K	14	0	2	12
SNP	2867266	rpoS	C->A	286	1	0	0	1

SNP	2867332	rpoS	C->A	220	1	0	0	1
DEL	2867337	rpoS	96 bp		1	0	0	1
MCDEL	2867337	rpoS	95 bp		1	0	0	1
SNP	2867369	rpoS	G->T	183	1	0	0	1
SNP	2867412	rpoS	G->T	140	1	0	1	0
SNP	2867415	rpoS	A->T	137	1	0	0	1
INS	2867427	rpoS	->T	coding (125/993 nt)	1	0	0	1
DEL	2867505	rpoS	1 bp		1	0	0	1
MOB	2867663	nlpD	IS1 -1 9 bp		1	0	0	1
SNP	2873568	cysC	C->A	R142L	1	0	0	1
SNP	2909317	pyrG	G->T	P117Q	1	0	0	1
SNP	2921667	gudP	G->T	A145E	1	0	0	1
SNP	2942497	gcvA	G->T	T24N	1	0	0	1
SNP	2951980	recD	G->T	A161E	1	0	0	1
SNP	2979375	lysR	C->A	R119S	1	0	1	0
SNP	3005386	ygeV	G->T	S134Y	1	0	1	0
SNP	3022843	xdhD	G->T	A510S	1	0	0	1
SNP	3028729	ygfS	G->T	A95E	1	0	0	1
SNP	3045425	ygfF	C->A	477	1	1	0	0
SNP	3052541	ubiH	G->A	978	1	0	0	1
SNP	3129632	yghO	G->T	pseudogene (512/1101 nt)	1	0	1	0
MCDEL	3130220	insH1	15273 bp		1	0	1	0
SNP	3134033	yghS/yghT	C->T	intergenic (-76/-98)	1	0	0	1
SNP	3160947	ygiQ	C->T	G67D	1	0	0	1
SNP	3163220	plsC	G->T	P88T	2	2	0	0
SNP	3214191	rpoD	C->A	A382E	1	0	0	1
SNP	3311086	pnp	G->T	T28N	1	0	1	0
SNP	3314317	infB	G->T	S566R	1	0	0	1
SNP	3415055	acrF	G->A	R8Q	1	0	0	1

MCDEL	3423697	rrfD/rrlD	884 bp		1	0	0	1
DEL	3476449	yheO/fkpA	1 bp		1	1	0	0
SNP	3486237	crp	C->A	L40M	1	0	0	1
SNP	3486285	crp	G->A	E56K	1	0	0	1
SNP	3486287	crp	G->T	E56D	1	0	1	0
INS	3486293	crp	->AAA	coding (174/633 nt)	1	0	1	0
SNP	3486307	crp	C->A	S63Y	1	0	0	1
SNP	3486541	crp	C->G	T141R	4	2	1	1
SNP	3486541	crp	C->A	T141K	3	0	1	2
SNP	3486552	crp	G->A	A145T	55	9	30	16
SNP	3486553	crp	C->A	A145E	34	1	4	29
SNP	3486630	crp	C->A	Q171K	4	0	2	2
SNP	3486642	crp	C->A	Q175K	2	0	0	2
SNP	3486688	crp	T->A	M190K	1	0	1	0
SNP	3486689	crp	G->A	M190I	1	0	0	1
SNP	3486699	crp	C->A	Q194K	1	0	1	0
SNP	3495553	nirB	C->A	H515N	1	0	0	1
SNP	3499065	cysG	C->A	A413E	1	0	1	0
SNP	3552550	malP/malT	C->A	intergenic (-77/-535)	1	0	0	1
SNP	3553012	malP/malT	G->T	intergenic (-539/-73)	1	0	0	1
SNP	3553301	malT	C->A	R73S	1	0	0	1
SNP	3553790	malT	G->T	A236S	1	0	0	1
SNP	3554017	malT	G->T	M311I	1	0	1	0
SNP	3554040	malT	G->T	C319F	1	0	1	0
SNP	3563467	glpD	G->T	1455	1	0	0	1
SNP	3607797	zntA	C->A	1347	1	0	1	0
MCDEL	3634818	yhiM	18363 bp		1	0	1	0
DEL	3634818	yhiMyhiS	17218 bp		1	0	1	0
MCDEL	3636128	yhiN	17019 bp		1	0	0	1

DEL	3636128	yhiNyhiS	15908 bp		1	0	0	1
SNP	3642775	rsmJ	G->T	P120T	1	0	1	0
SNP	3661457	mdtF	C->A	1044	1	1	0	0
SNP	3716661	yiaD	G->A	A39T	1	0	0	1
SNP	3843319	yicO	G->T	471	1	0	1	0
SNP	3901073	bglH	G->A	T383I	1	0	0	1
SNP	3923701	rsmG/mnmG	G->T	intergenic (-21/+43)	1	0	0	1
SNP	3973815	rffH	G->T	208	1	0	0	1
SNP	4014326	metE	C->T	A425V	1	0	0	1
SNP	4048946	polA	C->T	1981	1	0	0	1
SNP	4125061	priA	G->T	R584S	1	0	0	1
SNP	4153806	argE	G->A	P348S	1	0	0	1
SNP	4158347	argH/oxyR	C->A	intergenic (+124/-143)	1	0	0	1
SNP	4159202	oxyR	C->A	T238N	1	0	1	0
SNP	4159802	sthA	G->T	A330E	1	0	1	0
SNP	4185571	rpoC	G->T	K74N	1	0	0	1
SNP	4185590	rpoC	C->A	R81S	1	0	1	0
SNP	4200510	hupA	C->A	P77Q	1	0	1	0
MCDEL	4208121	rrsE/gltV/rrlE/rrfE	4977 bp		1	1	0	0
SNP	4242986	malG	G->T	531	1	0	1	0
SNP	4395275	nnr	C->A	R404S	1	0	0	1
SNP	4498115	intB/insC1	C->A	intergenic (+175/-157)	1	0	0	1
SNP	4535002	yjhR	G->T	pseudogene (212/213 nt)	1	0	0	1
SNP	4577881	mcrC	C->T	78	1	0	0	1
SNP	4580248	hsdS	G->T	1215	1	0	1	0
SNP	4640190	arcA	C->A	M39I	1	0	1	0

Paper II

A Nanobody:GFP bacterial platform that enables functional enzyme display and easy quantification of display capacity

Sofie Wendel, Emil C. Fischer, Virginia Martínez, Susanna Seppälä, Morten H.H. Nørholm

Abstract

Background: Bacterial surface display is an attractive technique for the production of cell-anchored, functional proteins and engineering of whole-cell catalysts. Although various outer membrane proteins have been used for surface display, an easy and versatile high-throughput-compatible assay for evaluating and developing surface display systems is missing.

Results: Using a single domain antibody (also called nanobody) with high affinity for GFP, we constructed a system that allows for fast, fluorescence-based detection of displayed proteins. The outer membrane hybrid protein LppOmpA and the autotransporter C-IgAP exposed the nanobody on the surface of *Escherichia coli* with very different efficiency. Both anchors were capable of functionally displaying the enzyme Chitinase A as a fusion with the nanobody, and this considerably increased expression levels compared to displaying the nanobody alone. We used flow cytometry to analyse display capability on single-cell versus population level and found that the signal peptide of the anchor has great effect on display efficiency.

Conclusions: We have developed an inexpensive and easy read-out assay for surface display using nanobody:GFP interactions. The assay is compatible with the most common fluorescence detection methods, including multi-well plate whole-cell fluorescence detection, SDS-PAGE in-gel fluorescence, microscopy and flow cytometry. We anticipate that the platform will facilitate future in-depth studies on the mechanism of protein transport to the surface of living cells, as well as the optimisation of applications in industrial biotech.

Keywords

surface display, nanobody, GFP, Chitinase A, LppOmpA, autotransporter, whole-cell catalysis

Background

Cell factories are a promising alternative to the problematic fossil fuel-based technologies currently employed in industry [1]. Cellular surface display of proteins is an attractive way to engineer whole-cell catalysts, thereby reducing time, cost and effort related to enzyme purification and one-time use of enzyme batches. Displaying proteins on the cell surface has the evident benefits of omitting any need to transport substrate or product across the cell wall, and may reduce toxicity effects due to the extracellular location of pathway components, substrates and products. The first successful cases of surface display were reported more than three decades ago [2, 3] but as pointed out by Schüürmann *et al.*, industrial development of whole-cell catalysts is lagging behind [4]. Surface display has been successfully carried out on several platforms such as yeast [5], phage [6] and bacteria [4] with several different cargos (typically antibodies or enzymes). Nevertheless, detailed fundamental understanding of the molecular mechanisms underlying surface display is lacking, thus complicating rational design [4, 7–9]. Furthermore, the development of display systems suffers from a lack of simple and wide-ranging assay methods, which would allow for easy detection and optimisation. Recently, an increasing number of studies have shown the use of fluorescently labelled antibodies as a non-generic method for visualising surface displayed proteins, and enabling not only their detection but also flow cytometric analysis and microscopy [10–12].

Several different membrane anchors have been explored for bacterial surface display; two of the main ones are autotransporters and outer membrane proteins from gram-negative bacteria (for review see e.g. [4, 13]). Autotransporters are typically involved in virulence and have been extensively explored for bacterial surface display [14–16]. Autotransporter proteins consist of an N-terminal signal peptide that

directs the polypeptide through the plasma membrane, a passenger domain that typically encodes a virulence factor, and a C-terminal translocation unit that enables the transport of the passenger domain across the outer membrane (fig 1a) [9]. While the transport of the passenger domain has given autotransporters their name, the whole protein as such is dependent on the barrel assembly machinery (BAM) complex to reach the cell surface [17, 18]. The C-terminal translocation unit of the *Neisseria gonorrhoeae* autotransporter IgA protease (C-IgAP) has been extensively characterised in terms of its mechanism of protein secretion as well as employed for surface display in *E. coli* [19, 20]. Native *E. coli* outer membrane proteins constitute a different class of surface display anchors. The LppOmpA fusion, consisting of the Lpp signal peptide followed by five transmembrane segments of Outer membrane protein A, has been successfully used to display enzymes such as hydrolases on the surface of *E. coli* (fig 1b) [11, 21].

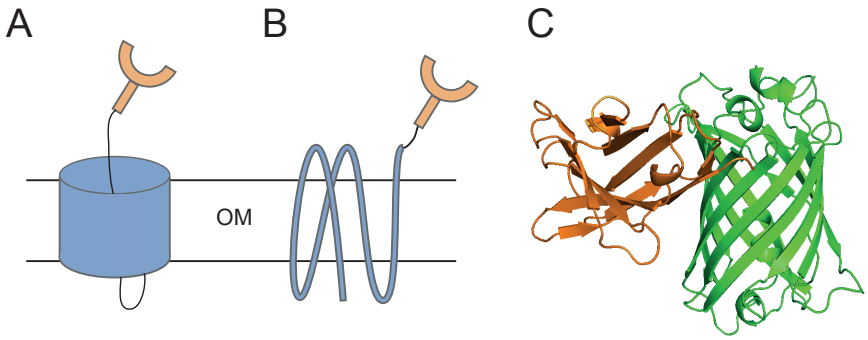


Figure 1. Illustrations of the nanobody:GFP complex and the outer membrane anchors. A, B. Schematic illustration of the nanobody (orange) **A**) as passenger of the autotransporter C-IgAP construct and **B**) fused to the outer membrane protein OmpA. **C.** Crystal structure of the enhancer nanobody binding GFP (PDB ID: 3OGO; [32]). OM: outer membrane.

Surface display of proteins has many parallels to bacterial membrane protein production, which is an inherently difficult process, dependent on proper balancing of the transport machineries and with optimal process conditions varying for different proteins [22–24]. The detection and optimisation of membrane protein production was dramatically simplified by the development of a GFP-fusion platform that enabled real-time monitoring, quantification and fast analysis of protein integrity and membrane association [25, 26]. As pointed out by Sun *et al.*, displaying GFP on the surface of cells could similarly be used to assess surface display levels [10]. This approach, however, comes with a major drawback: cells producing folded GFP would be fluorescent regardless of whether the GFP protein was actually displayed on the cell surface or not, i.e. if the protein remained in the cytoplasm or periplasm. Only with methods like advanced microscopy or complicated separation of compartments could one differentiate between surface localised and cyto/periplasmic GFP signal, and no conclusions about display efficiency could be drawn from a simple readout like whole-cell fluorescence. We therefore set out to develop a fluorescence-based method for surface display evaluation using an alternative approach, making use of a single-chain antibody molecule known as a nanobody.

Nanobodies (NB), found in camelids and sharks, are single domain antibodies that carry out the same function as full-size antibodies whilst consisting of only a variable heavy fragment [27, 28]. The small size of nanobodies makes them convenient protein tags, and the absence of essential disulphide bonds makes them easy to produce in *E. coli* [29, 30]. Kirchhofer *et al*. developed nanobodies that bind GFP with high specificity and affinity in a stable complex; in fact, the complex is stable enough to sustain denaturing SDS-PAGE

analysis (fig 1c) [31, 32]. Here, we have constructed a system for fluorescence- based detection of surface display by fusing the GFP-nanobody to different outer membrane anchors and visualising the displayed protein by adding purified GFP to whole cells.

Results

Construction of nanobody modules for surface display

GFP as reporter for surface displayed proteins is problematic, because it is difficult to differentiate between intracellular and surface displayed protein. Therefore, we used a complementary approach where the surface displayed protein is fused to a GFP-nanobody and subsequently detected using purified GFP added from the outside (fig 2a).

Two different display modules containing the nanobody were constructed, using the previously described GFP-enhancer-nanobody sequence [31]. As anchors, we chose two commonly used outer membrane proteins: We designed one display vector containing an Outer membrane protein A (OmpA) domain, and one vector containing an autotransporter domain, in both cases using the high-copy plasmid pKS1, herein called pK [33]. The outer membrane protein-based vector pK:LppOmpA-NB contains the N-terminal signal peptide of the *E. coli lpp* gene (residues M1 to Q29), followed by residues N66 to G180 of OmpA (forming five beta-strand transmembrane segments) and a C-terminally-fused nanobody sequence (fig 2b). An alternative vector, pK:pelB-NB-C-IgAP, was constructed by fusing the nanobody in-between the pelB signal peptide and the C-terminal domain of the *N. gonorrhoeae* autotransporter IgA protease (C- IgAP) (fig 2b). In both cases, protein production is under the control of the rhamnose-inducible *rhaPBAD* promoter.

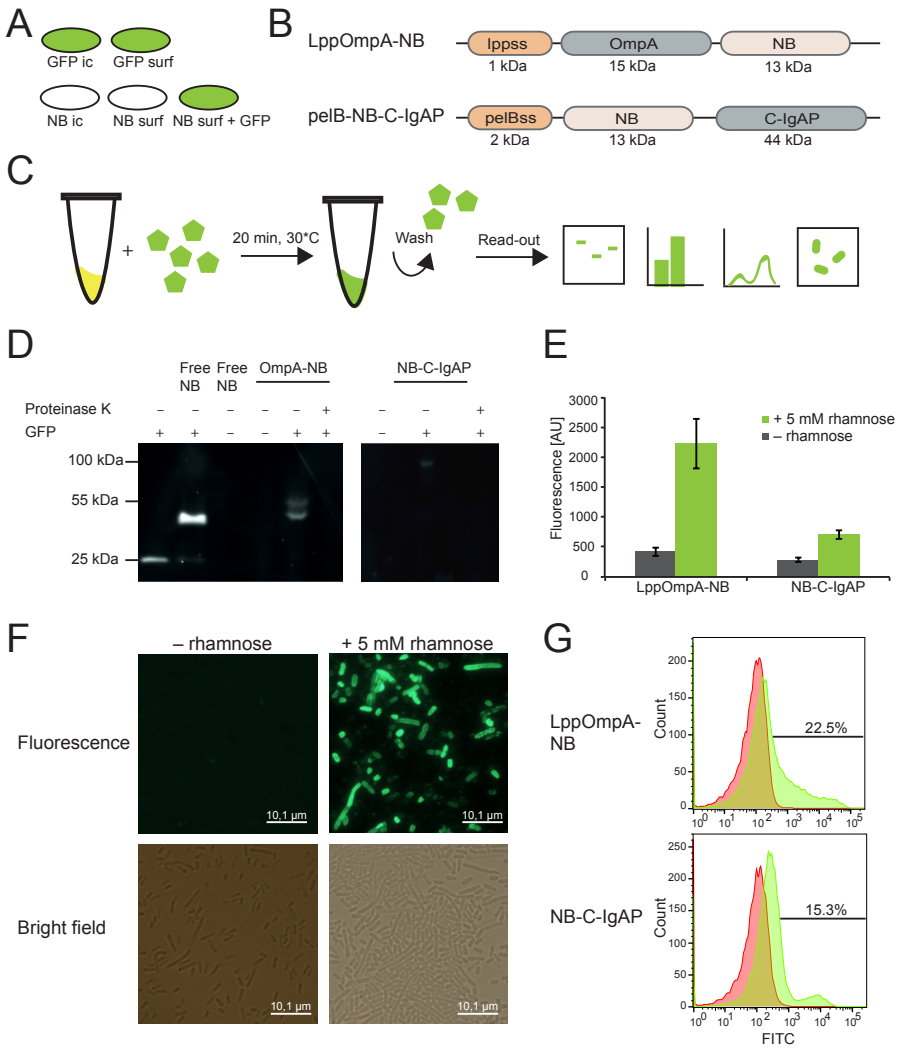


Figure 2. Characterisation of the NB:GFP platform. A. Illustration of the principal difference between displaying GFP and displaying the nanobody on the surface of the cell. GFP gives the cell a fluorescent glow whether produced intracellularly (ic) or on the surface (surf). In contrast, the only way a cell producing the nanobody can be fluorescent is if the nanobody is displayed on the surface and accessible to extracellular GFP. **B.** Protein schemes for the OmpA and autotransporter constructs. An N-terminal signal sequence (lppss and pelBss) precedes the OmpA anchor followed by the nanobody, or the nanobody followed by the C-IgAP anchor,

respectively. **C.** Workflow of GFP assay: cells producing the nanobody are incubated with free GFP; unbound GFP is washed off and the fluorescence signal from GFP bound to nanobody on cells is assayed using SDS-PAGE, fluorescence measurements and microscopy. **D.** In-gel fluorescence of purified GFP; purified GFP mixed with purified NB; NB; whole cells displaying OmpA-NB or NB-C-IgAP with and without GFP and with and without proteinase K treatment. The same amount of cells was loaded in each lane for whole cell samples. **E.** Whole-cell fluorescence measurement with and without rhamnose induction. Values are averages of three biological replicates and bars show standard error. **F.** Bright field and fluorescence microscopy images of OmpA-NB displayed on *E. coli* cells, with and without rhamnose induction. **G.** Flow cytometry profiles of pK:LppOmpA-NB and pK:NB-C-IgAP with induction (green) and without induction (red). Percentage numbers show the fraction of cells that are fluorescent.
Lppss: lpp signal sequence; pelBss: pelB signal sequence; OmpA: Outer membrane protein A; NB: nanobody; C-IgAP: C-terminal of IgA protease.

Functional, surface displayed nanobody is robustly assayed using GFP

pK:LppOmpA-NB and pK:NB-C-IgAP were transformed into *E. coli* BL21(DE3) and protein production was induced in liquid culture by the addition of 5 mM rhamnose. After 3 h of induction, cells were harvested, resuspended in buffer and incubated with purified GFP for 20 min at 30°C. Cells were harvested and washed twice with buffer to remove any unbound GFP; the repeated centrifugation steps also ensured that only whole cells were assayed. The washed cells were then subjected to (1) plate reader fluorescence measurement, (2) SDS-PAGE and in-gel fluorescence analysis, (3) flow cytometry analysis and (4) fluorescence microscopy (fig 2c). In all cases we could detect a fluorescence signal, showing the versatility of the nanobody:GFP platform (fig 2d-g).

Both the autotransporter C-IgAP and LppOmpA anchors successfully displayed the nanobody, as confirmed by in-gel fluorescence of OD-normalised whole-cell samples after incubation with GFP (fig 2d). The very fact that cells are fluorescing shows that the nanobody is accessible from the outside of the cell, and surface localisation is further confirmed by Proteinase K assay, removing all signal (fig 2d). The GFP signal is confined to bands corresponding to a complex of GFP bound to the NB construct (theoretical sizes 55 and 85 kDa, respectively) and none of the fluorescence appear to originate from free GFP (27 kDa, fig 2d). Whole cell fluorescence was measured in a plate reader and used to evaluate and quantify display ability of the entire bacterial population (fig 2e). Induced cultures were highly fluorescent compared to uninduced cultures, and the cultures containing the LppOmpA anchor showed approximately three times higher fluorescence than the C-IgAP cultures. Nanobody-displaying cells were also visualised using fluorescence microscopy: uninduced cells incubated with GFP and then washed prior to microscopy showed no fluorescence signal, while induced cells were strongly fluorescent (fig 2f). Single cell display behaviour was analysed by flow cytometry, which revealed that a fraction of the cells (22.5% for LppOmpA and 15.3% for C-IgAP) were responsible for the majority of the fluorescence (fig 2g). Also, the top LppOmpA expressers reached a 5-fold higher fluorescence value than the corresponding C-IgAP cells. C-IgAP producing cells were much more negatively affected by induction than LppOmpA cells; OD dropped by 70% when inducing NB-C-IgAP, and this was not affected by varying inducer concentration (Additional file 1: Fig S1). In contrast, the density of LppOmpA cultures decreased gradually and less dramatically upon induction (Additional file 1: Fig S1). Background fluorescence from the added GFP was essentially absent, as seen by

whole cell fluorescence for uninduced cells, in-gel fluorescence, microscopy, and flow cytometry (fig 2e-g). The detection system was functional and robust in both small and large format, with 96-well format allowing high-throughput analyses.

Displaying a functional enzyme as GFP-detectable nanobody fusions

The application of the nanobody platform was further tested by making sandwich fusions to the Chitinase A enzyme from *Serratia marcescens* (courtesy of Prof. Victor de Lorenzo, CNB, Madrid). ChiA is an industrially relevant enzyme for biotechnology applications [34]. The chitinase was fused either N- or C- terminally to the nanobody in the pK:LppOmpA-NB construct, and N-terminally to the nanobody in the pK:NB-C-IgAP construct, resulting in a total passenger size of 72 kDa (fig 3a). The proteins were subsequently produced in *E. coli* BL21(DE3) and surface exposure was assayed using GFP as described above. Successful display was confirmed by in-gel fluorescence (fig 3b), and whole cell fluorescence (fig 3c). This demonstrated that the nanobody was fully functional and binding its antigen GFP also when fused to another, large protein, and even when sandwiched in-between two proteins. As for the initial nanobody constructs, significant differences in display efficiency were observed as an effect of anchor usage. Interestingly, fusing the chitinase to the nanobody considerably increased display efficiency for both the LppOmpA-anchor and C-IgAP (fig 3d-e). Furthermore, the position of the nanobody in the construct influenced the surface presentation; pK:LppOmpA-NB-ChiA showed substantially higher fluorescence than pK:LppOmpA-ChiA-NB.

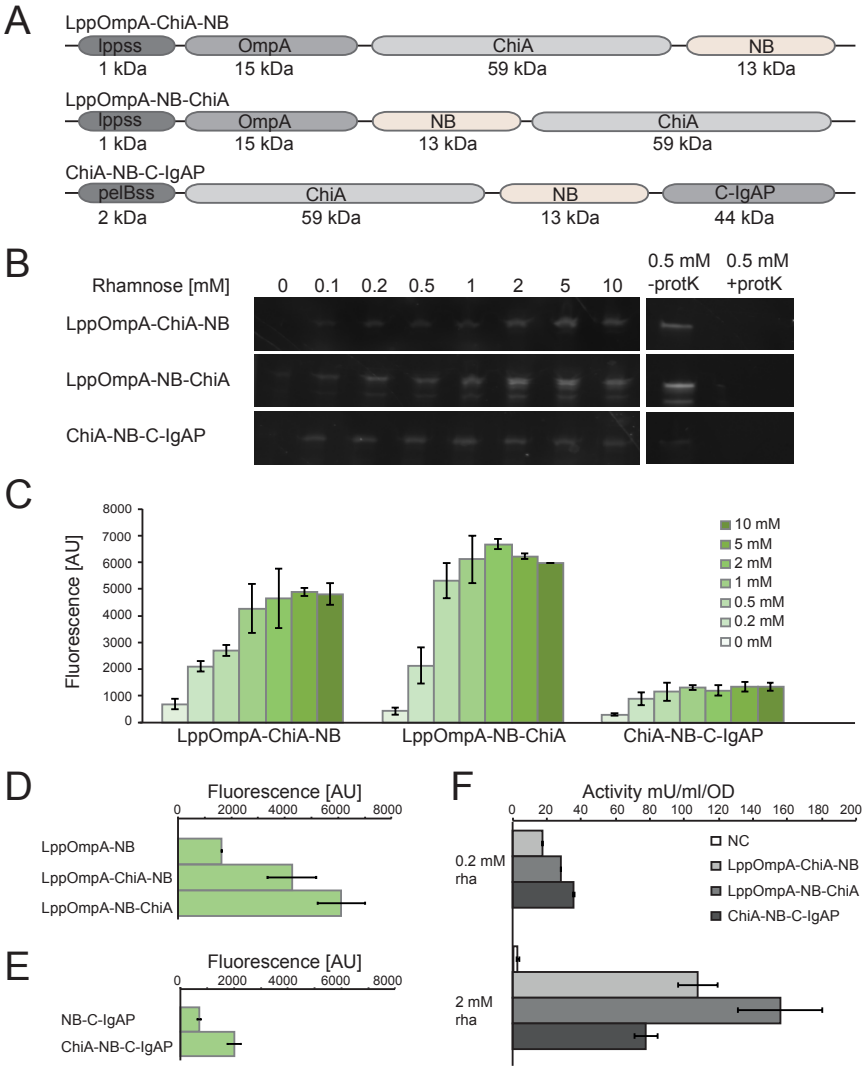


Figure 3. Functional display of Chitinase A using both display anchors. A. Protein schemes for Chitinase A-NB fusions. The ChiA protein was fused either in-between OmpA and NB, or C-terminally to LppOmpA-NB. With the C-IgAP anchor, ChiA was fused N-terminally to NB-C-IgAP. **B.** In-gel fluorescence of rhamnose titrations of chitinase-nanobody fusions, and in-gel fluorescence after addition of proteinase K. The same amount of cells was loaded in each lane. **C.** Whole cell fluorescence for rhamnose titration of

each of the chitinase-nanobody fusions. Values are averages of biological duplicates, error bars are standard errors. **D.** Whole cell fluorescence for LppOmpA constructs with and without ChiA, induced with 1 mM rhamnose. Values are averages of biological duplicates, bars show standard error. **E.** Whole cell fluorescence for C-IgAP constructs with and without ChiA, induced with 1 mM rhamnose. Values are averages of biological triplicates, bars show standard error. **F.** Specific chitinase activity for nanobody- chitinase fusions at two different inducer concentrations, normalised to OD. Values are averages of biological duplicates, bars are standard errors. Lppss: lpp signal sequence; pelBss: pelB signal sequence; OmpA: Outer membrane protein A; ChiA: Chitinase A; NB: nanobody; C-IgAP: C-terminal of IgA protease.

The rhamnose promoter is highly titratable [35, 36]. To test this tunability in our system, we varied inducer concentration from 0 to 10 mM rhamnose (fig 3b-c). Based on in-gel fluorescence analysis and plate reader data, increasing the concentration of rhamnose led to higher protein production, but the effect levelled off at higher concentrations. The LppOmpA constructs showed better tunability by rhamnose than the autotransporter version, which was largely unaffected by inducer concentration.

To confirm the functionality of the surface displayed enzyme, chitinolytic activity was assayed *in vivo*, for the same amount of cells, at two different inducer concentrations, 0.2 mM and 2 mM rhamnose (fig 3f). The chitinase was active with all surface display constructs, with higher activity for the LppOmpA fusions pK:LppOmpA-NB-ChiA (156 ± 25 mU/ml/OD at 2 mM rhamnose induction) and pK:LppOmpA- ChiA-NB (108 ± 11 mU/ml/OD) than for the autotransporter variant pK:ChiA-NB-C-IgAP (77 ± 7 mU/ml/OD). Activity levels correlated well with GFP-based expression data

measured with plate reader and in-gel fluorescence for the LppOmpA fusions, with a doubling of fluorescence corresponding to a doubling in activity (fig 3c and 3f). This correspondence between enzymatic activity and fluorescence data showed that the NB:GFP assay gives a reliable indication of how much functional protein is displayed on the cell surface for LppOmpA. For pK:ChiA-NB-C-IgAP the correlation is weaker, with activity levels varying more than fluorescence levels for the two inducer concentrations. Controls without chitinase showed no background activity.

Flow cytometry analysis reveals two populations of cells and confirms varying display efficiency

To study the display efficiency on a single-cell level, cells were analysed by flow cytometry (fig 2g and fig 4). This revealed two disparate populations of cells, both when using the LppOmpA anchor and the C- IgAP anchor (fig 4a and b, respectively), with only one of the populations presenting the protein fusion on the cell surface, as previously observed [10, 37]. The proportion of fluorescent cells varied from only 15.3% for pK:NB-C-IgAP to 41.9% for pK:LppOmpA-ChiA-NB (fig 4c). Interestingly, the relative amount of displayers was virtually identical for pK:LppOmpA-ChiA-NB (41.9%) and pK:LppOmpA- NB-ChiA (41.8%), but the mean fluorescence as measured by flow cytometry is 2.5-fold higher for pK:LppOmpA-NB-ChiA, in line with whole-cell fluorescence (fig 3c). Thus, the pK:LppOmpA-NB- ChiA population contained cells that had very high fluorescence per cell, while the fluorescent population was more homogeneous in the case of pK:LppOmpA-ChiA-NB. This highlights the different levels of information obtained from the different methods, and that high protein titers not necessarily correspond to high production per cell. The strikingly positive effect of Chitinase A on protein display levels was also evident with flow

cytometry, with the proportion of fluorescent cells increasing by up to 86% when adding Chitinase A to the LppOmpA fusion (compare fig 2g and 4a), and by 103% for the corresponding C-IgAP fusion proteins (compare fig 2g and 4b). Under control of the P_{trc} promoter, the LppOmpA-NB construct formed a single population (Additional file 2: Fig S2).

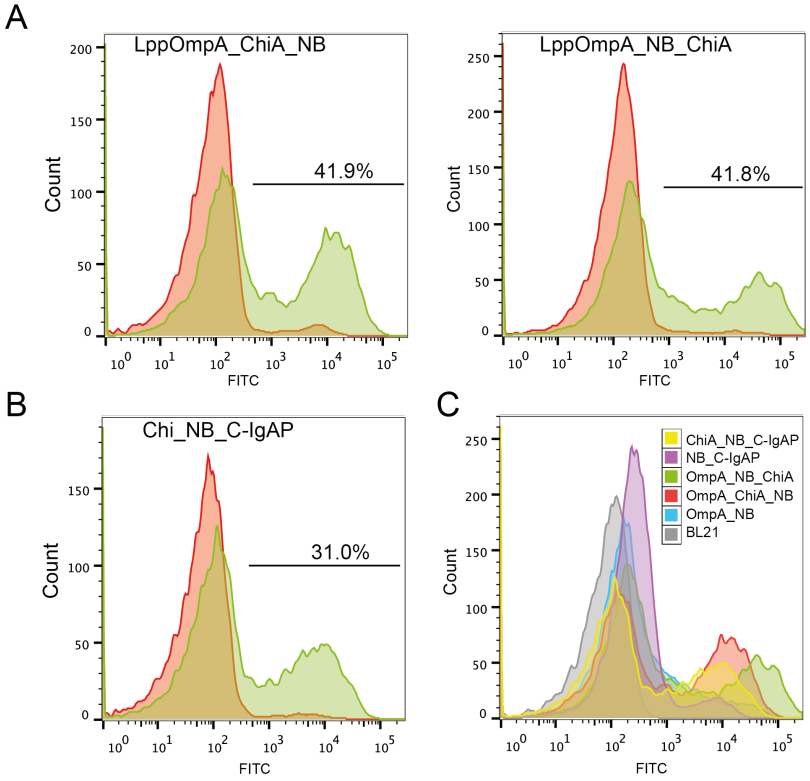


Figure 4. FACS-analysis of Chitinase A-nanobody surface display fusions. Fluorescence was measured by flow cytometry 3 h after rhamnose induction of cells in exponential growth followed by 20 min incubation with purified GFP, and two steps of washing. **A.** Chitinase-nanobody fusions surface displayed with LppOmpA, with (green) and without (red) rhamnose induction. **B.** Chitinase-nanobody fusion displayed with C-IgAP, with (green) and without (red) rhamnose induction. **C.** Overlays of the flow cytometry profiles for all fusions, with rhamnose induction: LppOmpA-NB,

LppOmpA-ChiA-NB, LppOmpA-NB- ChiA, NB-C-IgAP, and ChiA-NB-C-IgAP.

Nanobody:GFP platform allows systematic study of signal peptide effects on surface display

With the assay running in a convenient 96-well format, it is possible to study how different parameters affect surface display in a high-throughput manner. As proof of concept, we studied the effect of the signal sequence on display efficiency. The challenging-to-display ChiA-NB-C-IgAP fusion was cloned into a set of nine low-copy pD881 plasmids, identical except for harbouring different signal peptides responsible for targeting the protein to the periplasm. We subsequently surveyed protein display using the NB:GFP assay, and compared whole-cell fluorescence values for the different constructs. Large variation was observed between the different signal sequences (fig 5a). Several of the signal peptides showed very low display capacity, in particular torA for which induced cells were negligibly fluorescent. The signal peptides dsbA, ompA, ompC, ompT, pelB, sufI and torT gave higher fluorescence values, whereas glll resulted in approximately twice as high fluorescence as the other signal peptides. Notably, pelB is the signal peptide used in the original pK construct, but, based on this result, it is only mediocre in comparison with glll. Therefore we followed up on this experiment by replacing the pelB signal peptide in the pK backbone with the gIII signal peptide and analysed protein production. In this high copy vector, the advantage of gIII over pelB disappeared (fig 5b). A signal sequence library was likewise constructed for the LppOmpA-ChiA-NB fusion. Again, the various signal sequences resulted in variation of enzyme display but without any exceptional high-displayer (Additional file 3: Fig S3).

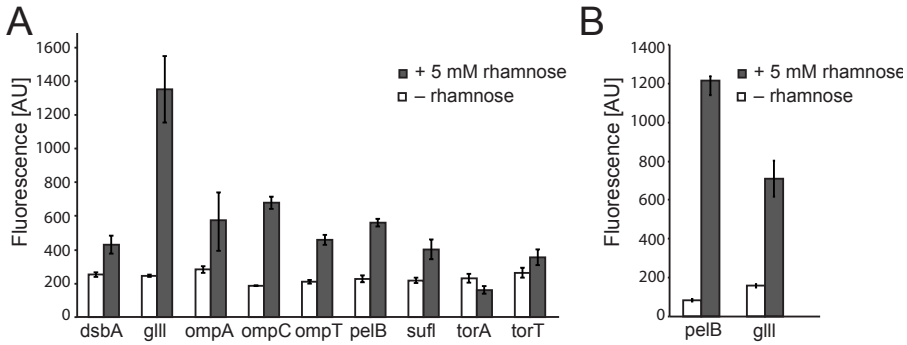


Figure 5. **Evaluation of a signal sequence library on display levels of nanobody-fused chitinase (ChiA-NB-C- IgAP). A.** The ChiA-NB-C-IgAP fusion was cloned into a set of nine vectors containing different signal peptides for directing the polypeptide to the periplasm. Fluorescence values were measured in a plate reader. Significant variation among the different peptides was observed, with gIII showing the highest signal. Values are averages of biological triplicates, bars standard error. **B.** Comparison of gIII and pelB signal peptides in the pK backbone. Values are averages of four biological replicates and bars show the standard error.

Discussion

Bacterial surface display of enzymatically active proteins is a promising strategy to engineer whole-cell catalysts, enabling simplification of production procedures as well as downstream processes. Even though many surface display systems have been successfully employed since the first bacterial anchors were developed in the 1980s (reviewed in [4]), unpredictable, cargo-dependent effects hinder rational design and optimization, as exemplified by Nicolay *et al*. who reported failure to display a number of passenger domains with a previously characterised autotransporter [8].

In this study we used a camelid-derived nanobody to construct a quantitative, inexpensive, and robust assay that allows for easy GFP-based screening and optimization of surface display systems in a high- throughput manner. By making use of the tight association between a surface-presented nanobody and externally applied GFP, we addressed a need for a fluorescence-based assay for surface display, underscored by the avid use of two-step antibody labelling procedures in the literature [10–12]. Furthermore, we demonstrate that the NB:GFP platform is compatible with all techniques commonly associated with fluorescent tags: whole cell fluorescence measurements, in-gel fluorescence analysis, flow cytometry, and microscopy. Previously established techniques for fluorescence-based evaluation of surface display include the use of small peptide tags, such as FLAG and myc [38], and of domains of staphylococcal protein A and streptococcal protein G [39], followed by detection with antibodies. In relation to these, the NB:GFP platform differs especially in that it is a one-step procedure enabling all down-stream analyses, in particular simplifying analysis of samples on protein gels. Furthermore, costs for especially monoclonal antibodies are high, whereas both NB and GFP are easily produced in *E. coli*.

Both of the two anchors used, LppOmpA and C-IgAP, enabled functional display of the nanobody, although the efficiency was higher with LppOmpA in terms of total protein production as well as display efficiency across a population of cells. The basis for the difference in fluorescence could alternatively be low expression of the C-IgAP construct gene, poor folding of the nanobody, or poor GFP-NB interaction at the surface. The observed fitness cost for cells induced for NB-C-IgAP production, however, indicates that the explanation is meagre protein display, and points to LppOmpA as a more robust anchor in our system. While autotransporters have been

reported as successful display anchors in many cases, our results are consistent with several studies where negative effects of autotransporter anchors were observed, with regards to cell viability [40], membrane integrity [41] and ultimately on surface display efficacy [8].

Although nanobodies are pharmaceutically interesting display targets in their own right (recently displayed in Gram positives [42] and *E. coli* [43], and routinely selected through phage-display [44]), our goal was to use the nanobody as a molecular biology tool for detection of other passenger-protein fusions. We applied the NB:GFP platform for surface display of the industrially relevant enzyme Chitinase A from *S. marcescens* [45, 46]. Its substrate chitin is one of the most abundant biomasses on Earth and degradation of chitin is attractive for e.g. bioethanol production, production of new materials and in the food industry [34].

Importantly, the nanobody was readily binding GFP independent of its position in the fusion protein: when displayed immediately linked to outer membrane proteins close to the cell surface; when being placed further away from the surface as fusions to the functional Chitinase A enzyme domain; and when being sandwiched in between the cell anchor and Chitinase A. This suggests that the nanobody is a robust fusion partner, also suitable for more complex designs such as multi-enzyme pathways. While GFP itself can be displayed on the cell surface [10, 47, 48], the complementary approach of displaying the nanobody and detecting it with externally added GFP circumvents the problem of false positives, since only binding with nanobody presented at the cell surface will yield a fluorescent signal.

Interestingly, adding Chitinase A to our protein fusions markedly enhanced display levels in spite of more than doubling the size of the fusion protein. A possible alternative explanation of the observed increase in fluorescence is that the nanobody became more accessible to GFP when fused to ChiA - this, however, seems unlikely given that the nanobody is sandwiched in-between two proteins in the best performing constructs. The apparent correlation between chitinolytic activity and GFP signal further supports an actual increased display level. Interactions between a passenger domain and the cell membrane have been suggested to be an important factor for translocation, and one could speculate that ChiA similarly stimulates translocation of its fusion partner [49]. In general, saturation of the transport machinery responsible for translocating proteins to the cell surface is a major bottleneck when overexpressing membrane protein genes [22]. Since Chitinase A is naturally secreted in its native gram-negative host, in contrast to the nanobody, it is likely that the biochemical properties of the protein are well suited for translocation in *E. coli* as well. The chitinase was active in all fusion combinations, and activity levels corresponded well with fluorescence data for LppOmpA, whereas correlation was weaker for the autotransporter version. This may be explained by some of the protein being halted in the periplasm, where the assay substrate might be accessible. This is in line with the observed fitness cost of NB-C-IgAP production. Although varying inducer concentrations made it possible to tune display to a certain degree, the increase in display eventually levelled off for the LppOmpA fusions, and C-IgAP display was only marginally affected. This fits well with previous reports that translocation is a generally limiting step in surface display [22, 47, 48].

The versatility of the NB:GFP platform allowed us to study the population behaviour by flow cytometric analyses. The LppOmpA fusions with Chitinase A constituted either a relatively homogeneous population of medium-expressers (fig 4a, left panel), or a slightly more heterogeneous population with some very highly expressing cells (fig 4a, right panel). This demonstrates how the evaluation of production efficiency may vary dependent on the read-out method, and how the versatility of the NB:GFP platform allows for optimisation of several important parameters. This will likely be of high value for development of robust whole-cell catalysts. The occurrence of two populations can in the context of the present study be attributed to the rhamnose promoter: when expressing the LppOmpA-NB construct under control of the IPTG-inducible P_{trc} promoter, we observed an almost homogeneous population distribution (Additional file 2: Fig S2).

We have used the platform to evaluate one of the key parameters for secretion, the N-terminal signal sequence responsible for targeting the nascent polypeptide to the periplasm prior to its translocation across the outer membrane [50]. This demonstrated the potential of the NB:GFP platform to study the effect of different signal peptides in surface presentation, in a multi-format setup. Fluorescence signals varied more than 6-fold between the lowest and highest displayed constructs. In consistence with previous reports [51], this highlights the importance of systematic process optimisation for surface display, and shows the feasibility of the developed technology. The negligibly low torA signal is not unexpected since this peptide only directs export of fully folded protein [52]. Compared to the original pK construct, all tested signal sequences gave a lower fluorescence signal, which is likely explained by the fact that pK is a high-copy plasmid whereas the signal peptide library backbone is a low-copy

vector. Inserted into pK, however, gIII no longer excelled compared to pelB. This shows that surface display is dependent on an intricate mesh of mechanisms, where plasmid copy number and signal sequence are two important, interdependent parameters, making display optimisation a complex process.

The NB:GFP system is a rapid assay for quantitative assessment of surface display, making use of GFP that is easily produced and purified in *E. coli*. The platform has the potential to ease systematic studies of surface display systems and drive quick optimisation of individual display systems in a multi-well format – however, optimisation and testing of the system for each individual protein will likely be needed. Furthermore, our hope is that the NB:GFP platform will facilitate the fundamental understanding of the molecular mechanisms behind the biogenesis of *E. coli* outer membrane proteins that can enable rational development of bacterial surface display systems and robust whole-cell biocatalysts in the future.

Conclusions

We developed an inexpensive, robust and quantitative surface display platform that allows for functional display of enzymes. Furthermore, we employed the nanobody:GFP platform for (1) expression of the industrially relevant enzyme Chitinase A and (2) evaluation of a signal peptide library's effect on surface display through easy quantitative screening.

Methods

Gene and vector design

All cloning was done with USER fusion (as described in Cavaleiro *et al*. [53]) unless otherwise specified, into the high-copy plasmid vector

pK (described in [33]) and is outlined in Additional file 4: Fig S4. C-IgAP sequence and the N-terminal signal peptide pelB were obtained from the de Lorenzo lab [19] and cloned into pK using oligos 525, 526, 527 and 528, forming pK:C-IgAP. pK:lpp-OmpA was constructed in several steps: Lpp and OmpA were amplified from *E. coli* K12 MG1655 genomic DNA with oligos Lpp-F and Lpp-OmpA-R, and OmpAR and Lpp-OmpA-F, respectively, thereby constructing the previously described LppOmpA chimera [21]. They were cloned into plasmid pGFP (described in [54]) using oligos pGFP_1 and pGFP_2, creating plasmid pLppOmpA-GFP. Later, LppOmpA-GFP was transferred to the pK vector with oligos GFPgenR and Lpp-F for LppOmpA-GFP amplification, and oligos 525 and 526 for opening pK. Subsequently, the nanobody sequence was ordered as a G-block (Genscript) codon optimized for *E. coli*, and cloned into plasmid pGFP with restriction cloning using *XhoI* and *HindIII* sites. The nanobody was then cloned C-terminally to LppOmpA, replacing GFP, using oligos 855, 856, 857 and 858, creating plasmid pK:LppOmpA-NB. N-terminal fusion of the nanobody to the autotransporter C-IgAP, forming pK:NB-C-IgAP, was done using oligos 1715, 1716, 1717, and 1718. The Chitinase A gene was provided by the de Lorenzo lab, and was inserted C- or N-terminally to LppOmpA by cloning with oligos 1889, 1890, 1891, 1892, 1898, and 1899. pK_ChiA-NB-AT was constructed by cloning with oligos 1889, 1895, 1898, and 1899. All oligos and plasmids can be found in Additional file 5: Table S1 and Additional file 6: Table S2, respectively.

Bacterial strains and culture conditions

Escherichia coli NEB5alpha strain (New England Biolabs) was used for cloning purposes and *E. coli* OneShot BL21(DE3) (ThermoScientific) were used for protein production and display. All cultures were

grown at 37°C in Luria Bertoni broth (LB) under agitation, unless otherwise noted, with kanamycin supplemented to 50 μg/ml to maintain the pK plasmids.

Production of surface displayed proteins

Surface display ORFs were under the control of the *rhaPBAD* promoter. *E. coli* BL21(DE3) cells containing the pK surface display plasmids were inoculated from overnight culture to OD600 0.1 in LB and grown at 37°C, 250-300 rpm, to an OD600 of 0.3-0.5, when protein production was induced. Induction was carried out using various concentration of rhamnose and expression of plasmid-encoded genes was allowed for 3 h at 30°C.

GFP and NB production and purification

His-tagged nanobody and Folding reporter GFP [55] under control of the T7 promoter were produced in *E. coli* SHuffle (New England Biolabs) and *E. coli* BL21 (DE3), respectively, by induction with 0.4 mM IPTG for 5 h. Cells were then resuspended in IMAC wash buffer (50 mM Tris-HCl, 10 mM imidazole, 500 mM NaCl, 10% glycerol [pH 7.5]), lysed by three passes through an EmulsiFlex-C5 homogenizer (Avestin) at 10 000–15 000 psi and any debris and unbroken cells were removed by centrifuging at 18 000 g at 4°C for 15 min. The supernatant containing the proteins of interest was loaded onto nickel- nitrilotriacetic acid (Ni2+-NTA) resin columns (HisTRAP) on an Äkta Pure system connected to an F9-C fraction collector (GE). The bound protein was washed extensively with IMAC wash buffer and was subsequently eluted by increasing the imidazole concentrations to 500 mM in a single step. The fractions containing the protein of interest were pooled and stored at –80°C for nanobody, and –20°C for GFP, until use.

Nanobody:GFP assay

Protein production was stopped by pelleting cells via centrifugation for 4 min at 2272 g in a ThermoScientific Multifuge X3 FR centrifuge. Cells were then resuspended in 50 µl 50 mM Tris buffer, and mixed with 50 µl 0.12 mg/ml GFP (final concentration 0.06 mg/ml). Cells and GFP were incubated for 20 min at 30°C at 250-300 rpm. Incubation was stopped by centrifugation of cells, 4 min at 2272 g. Cells were washed twice with 300 µl 50 mM Tris buffer, and then resuspended in 220 µl 50 mM Tris buffer before downstream analyses.

GFP signal detection and OD measurement using plate reader

200 µl cell suspension was transferred to an opaque microtiter plate (Sigma-Aldrich) and GFP signal was read in a SynergyMx plate reader (BioTek) at gain 80. 40 µl cells were then transferred to 160 µl 50 mM Tris buffer (1/5x dilution) in a transparent microplate (Greiner Bio-One) and optical density at 600 nm was measured.

SDS-PAGE analysis

Cells were resuspended to a concentration of 0.05 ODU/µl and 10 µl were mixed with 5 µl 2x Laemmli sample buffer and 0.5 µl benzonase nuclease (≥250 units/µl, Sigma), after which the whole sample was loaded onto a 4-20% Mini-PROTEAN® TGXTM gel (Bio-Rad) and run for 35 min at 150V. Fluorescent protein bands were visualised with the G:Box bioimager (Syngene) using UV-light filter, and total protein was assessed by staining with InstantBlue (Expedeon).

Fluorescence microscopy

3 µl cells were pipetted onto Poly-prep microscopy slides (Sigma) and studied in a Leica DM4000B fluorescence microscope at 100x magnification, using Leica Application Suite v4.0 for capturing

images. The GFP fluorophore was excited and signal detected using an excitation filter with bandpass 470/40 and a suppression filter with bandpass 525/50.

Proteinase K accessibility assay

Cells were harvested and resuspended in 50 µl PBS buffer before adding 1.5 µl Proteinase K (final concentration 0.58 mg/ml) (ThermoScientific). Samples were incubated for 30 min at 37°C, after which the accessibility was assessed by carrying out the NB:GFP assay and analysing samples on SDS-PAGE as described above, starting with a centrifugation and washing step to remove all cleaved off protein.

Flow cytometry

Flow cytometry measurements were performed on a FACS Aria (Becton Dickinson, San Jose, USA) with 488 nm excitation from a blue solid-state laser. Cells were diluted 1:100 in PBS for analysis. At least 20,000 cells were collected for each measurement. FlowJo (Treestar) was used for data analysis**.**

Chitinase activity assay

Chitinase activity was analysed using the Chitinase Assay Kit (Sigma-Aldrich) and carried out according to the manufacturer's instructions. Briefly, cells induced for chitinase production during 4 h were resuspended to 0.0044 OD600 units/µl and 10 µl (0.044 OD600 units) were mixed with 10 µl 4-Nitrophenyl N,N'-diacetyl-β-D-chitobioside (1 mg/ml). Samples were incubated at 37°C for 30 min, and the reaction was stopped by addition of 200 µl 39 mM sodium carbonate solution. Cells were harvested, and 200 µl supernatant was transferred to a microtiter plate (Greiner Bio-One) and the concentration of *p*-nitrophenol was measured as absorbance at 405

nm in a SynergyMx plate reader (BioTek). Specific activity was calculated as absorption per time per volume against a *p*-nitrophenol standard.

Cloning of P_{trc} construct

The open reading frame from plasmid pK:LppOmpA-NB was amplified using primers 2148 and 2152 and USER cloned into the linearized backbone pCDF_sl3m:Ptrc made from amplifying plasmid pCDF_sl3m:Ptrc-GFP [56] with oligos 2155 and 2197 using the high-fidelity polymerase Phusion U (ThermoScientific). Expression and detection was carried out as described in the methods section apart from: (1) using 1 mM IPTG as final inducer concentration instead of L-Rhamnose and (2) supplementing LB growth medium with a final concentration of 50 μg/ml spectinomycin instead of kanamycin.

Signal sequence library cloning

A panel of signal sequences (dsbA, gIII, ompA, ompC, ompT, pelB, sufI, torA, torT) in linearized vector pD881 was acquired from DNA2.0. Cloning of surface display constructs was done in close agreement with manufacturer's instructions by treatment with *LguI* and T4 DNA Ligase (Thermo Scientific) in Tango buffer supplemented with ATP (Thermo Scientific) prior to transformation into NEB5alpha (New England Biolabs). Oligos 2236 and 2334 were used for ChiA-NB-C-IgAP library and oligos 2235 and 2539 for the LppOmpA-ChiA-NB ditto. The gIII signal sequence was subsequently cloned to replace the pelB peptide in pK:ChiA-NB-C-IgAP using oligos 2647 and 2648.

Declarations

Availability of data and material

All material available upon request.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by The Novo Nordisk Foundation and a PhD grant from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme [FP7-People-2012-ITN], under grant agreement No. 317058, "BACTORY". SS is the recipient of VILLUM Foundation's Young Investigator Programme grant VKR023128. ECF is recipient of the Novo Scholarship Programme.

Authors' contributions

SW carried out cloning of the nanobody and chitinase constructs, development of the assay, arranged and conducted the main part of experiments. ECF cloned the signal sequence libraries, the P_{trc} construct, and assisted with experiments. VM carried out flow cytometry analyses and helped interpret the data. SS developed the experimental rational, carried out cloning and experimental design. MN guided the design of the study and cloned the pK:LppOmpA construct. All authors assisted with writing the manuscript.

Acknowledgements

We thank Dr. Stefan Kol for purification of GFP and nanobody. We are thankful to Prof. Victor de Lorenzo for plasmids encoding C-IgAP

and Chitinase A. SW is recipient of a doctoral fellowship grant from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme [FP7-People-2012-ITN], under grant agreement No. 317058, "BACTORY". SS is the recipient of VILLUM Foundation's Young Investigator Programme grant VKR023128. ECF is recipient of the Novo Scholarship Programme.

References

1. Lee SY, Mattanovich D, Villaverde A: **Systems metabolic engineering, industrial biotechnology and microbial cell factories.** *Microb Cell Fact* 2012, **11**:156.

2. Freudl R, MacIntyre S, Degen M, Henning U: **Cell surface exposure of the outer membrane protein OmpA of *Escherichia coli* K-12**. *J Mol Biol* 1986, **188**:491–494.

3. Charbit A, Boulain JC, Ryter A, Hofnung M: **Probing the topology of a bacterial membrane protein by genetic insertion of a foreign epitope; expression at the cell surface.** *EMBO J* 1986, **5**:3029–37.

4. Schüürmann J, Quehl P, Festel G, Jose J: **Bacterial whole-cell biocatalysts by surface display of enzymes: toward industrial application.** *Appl Microbiol Biotechnol* 2014, **98**:8031–46.

5. Chao G, Lau WL, Hackel BJ, Sazinsky SL, Lippow SM, Wittrup KD: **Isolating and engineering human antibodies using yeast surface display**. *Nat Protoc* 2006, **1**:755–768.

6. Bradbury AR., Marks JD: **Antibodies from phage antibody libraries**. *J Immunol Methods* 2004, **290**:29–49.

7. Veiga E, De Lorenzo V, Fernández LA: **Structural tolerance of bacterial autotransporters for folded passenger protein domains**. *Mol Microbiol* 2004, **52**:1069–1080.

8. Nicolay T, Lemoine L, Lievens E, Balzarini S, Vanderleyden J, Spaepen S: **Probing the applicability of autotransporter based surface display with the EstA autotransporter of *Pseudomonas stutzeri* A15.** *Microb Cell Fact* 2012, **11**:158.

9. van Ulsen P, Rahman SU, Jong WSP, Daleke-Schermerhorn MH, Luirink J: **Type V secretion: from biogenesis to biotechnology.** *Biochim Biophys Acta* 2014, **1843**:1592–611.

10. Sun F, Pang X, Xie T, Zhai Y, Wang G, Sun F: **BrkAutoDisplay: functional display of multiple exogenous proteins on the surface of *Escherichia coli* by using BrkA autotransporter**. *Microb Cell Fact* 2015, **14**:129.

11. Liu R, Yang C, Xu Y, Xu P, Jiang H, Qiao C: **Development of a whole-cell biocatalyst/biosensor by display of multiple heterologous proteins on**

the *Escherichia coli* cell surface for the detoxification and detection of organophosphates. *J Agric Food Chem* 2013, **61**:7810–7816.

12. Park TJ, Heo NS, Yim SS, Park JH, Jeong KJ, Lee SY: **Surface display of recombinant proteins on *Escherichia coli* by BclA exosporium of Bacillus anthracis.** *Microb Cell Fact* 2013, **12**:81.

13. van Bloois E, Winter RT, Kolmar H, Fraaije MW: **Decorating microbes: surface display of proteins on *Escherichia coli*.** *Trends Biotechnol* 2011, **29**:79–86.

14. Jose J, Bernhardt R, Hannemann F: **Cellular surface display of dimeric Adx and whole cell P450- mediated steroid synthesis on *E. coli***. *J Biotechnol* 2002, **95**:257–268.

15. Schumacher SD, Hannemann F, Teese MG, Bernhardt R, Jose J: **Autodisplay of functional CYP106A2 in *Escherichia coli*.** *J Biotechnol* 2012, **161**:104–12.

16. Kranen E, Detzel C, Weber T, Jose J: **Autodisplay for the co-expression of lipase and foldase on the surface of *E. coli*: washing with designer bugs.** *Microb Cell Fact* 2014, **13**:19.

17. Leyton DL, Rossiter AE, Henderson IR: **From self sufficiency to dependence: mechanisms and factors important for autotransporter biogenesis.** *Nat Rev Microbiol* 2012, **10**:213–25.

18. Roman-Hernandez G, Peterson JH, Bernstein HD: **Reconstitution of bacterial autotransporter assembly using purified components.** *Elife* 2014, **3**:e04234.

19. Veiga E, Sugawara E, Lorenzo Â De, Nikaido H, Ferna LA: **Export of autotransported proteins proceeds through an oligomeric ring shaped by C-terminal domains**. *EMBO J* 2002, **21**:2122–2131.

20. Veiga E, Lorenzo V De, Ferna LA: **Autotransporters as Scaffolds for Novel Bacterial Adhesins: Surface Properties of *Escherichia coli* Cells Displaying Jun/Fos Dimerization Domains**. *J Bacteriol* 2003, **185**:5585–5590.

21. Earhart CF: **Use of an Lpp-OmpA Fusion Vehicle for Bacterial Surface Display**. *Hybrid proteins* 2000, **326**:506–516.

22. Wagner S, Baars L, Ytterberg a J, Klussmeier A, Wagner CS, Nord O, Nygren P-A, van Wijk KJ, de Gier J-W: **Consequences of membrane protein overexpression in *Escherichia coli*.** *Mol Cell proteomics* 2007, **6**:1527–1550.

23. Wagner S, Klepsch MM, Schlegel S, Appel A, Draheim R, Tarry M, Wijk KJ Van, Slotboom DJ, Persson JO, Gier J De: **Tuning *Escherichia coli* for membrane protein overexpression**. *Proc Natl Acad Sci* 2008, **105**:14371–14376.

24. Schlegel S, Hjelm A, Baumgarten T, Vikström D, Gier J De: **Bacterial-based membrane protein production**. *Biochim Biophys Acta - Mol Cell Res* 2013, **1843**:1739–1749.

25. Daley DO, Rapp M, Granseth E, Melén K, Drew D, von Heijne G: **Global topology analysis of the *Escherichia coli* inner membrane proteome.** *Science* 2005, **308**:1321–3.

26. Sonoda Y, Newstead S, Hu N-J, Alguel Y, Nji E, Beis K, Yashiro S, Lee C, Leung J, Cameron AD, Byrne B, Iwata S, Drew D: **Benchmarking membrane protein detergent stability for improving throughput of high-resolution X-ray structures.** *Structure* 2011, **19**:17–25.

27. Muyldermans S, Baral TN, Retamozzo VC, De Baetselier P, De Genst E, Kinne J, Leonhardt H, Magez S, Nguyen VK, Revets H, Rothbauer U, Stijlemans B, Tillib S, Wernery U, Wyns L, Hassanzadeh-Ghassabeh G, Saerens D: **Camelid immunoglobulins and nanobody technology**. *Vet Immunol Immunopathol* 2009, **128**:178–183.

28. Wesolowski J, Alzogaray V, Reyelt J, Unger M, Juarez K, Urrutia M, Cauerhff A, Danquah W, Rissiek B, Scheuplein F, Schwarz N, Adriouch S, Boyer O, Seman M, Licea A, Serreze D V., Goldbaum FA, Haag F, Koch-Nolte F: **Single domain antibodies: Promising experimental and therapeutic tools in infection and immunity**. *Med Microbiol Immunol* 2009, **198**:157–174.

29. Govaert J, Pellis M, Deschacht N, Vincke C, Conrath K, Muyldermans S, Saerens D: **Dual Beneficial Effect of Interloop Disulfide Bond for Single Domain Antibody Fragments**. *J Biol Chem* 2012, **287**:1970–1979.

30. Olichon A, Surrey T: **Selection of Genetically Encoded Fluorescent Single Domain Antibodies Engineered for Efficient Expression in *Escherichia coli***. *J Biol Chem* 2007, **282**:36314–36320.

31. Kirchhofer A, Helma J, Schmidthals K, Frauer C, Cui S, Karcher A, Pellis M, Muyldermans S, Casas- delucchi CS, Cardoso MC, Leonhardt H, Hopfner K, Rothbauer U: **Modulation of protein properties in living cells using nanobodies**. *Nat Struct Mol Biol* 2009, **17**:133–138.

32. Kubala MH, Kovtun O, Alexandrov K, Collins BM: **Structural and thermodynamic analysis of the GFP:GFP-nanobody complex.** *Protein Sci* 2010, **19**:2389–401.

33. Söderström B, Skoog K, Blom H, Weiss DS, von Heijne G, Daley DO: **Disassembly of the divisome in *Escherichia coli*: Evidence that FtsZ dissociates before compartmentalization**. *Mol Microbiol* 2014, **92**:1–9.

34. Martínez JP, Falomir MP, Gozalbo D: **Chitin: A Structural Biopolysaccharide with Multiple Applications**. In *eLS. John Wiley & Sons, Ltd: Chichester.*; 2014.

35. Haldimann A, Daniels LL, Wanner BL: **Use of new methods for construction of tightly regulated arabinose and rhamnose promoter fusions in studies of the *Escherichia coli* phosphate regulon**. *J Bacteriol* 1998, **180**:1277–1286.

36. Giacalone MJ, Gentile AM, Lovitt BT, Berkley NL, Gunderson CW, Surber MW: **Toxic protein expression in *Escherichia coli* using a rhamnose-based tightly regulated and tunable promoter system**. *Biotechniques* 2006, **40**:355–364.

37. Yang X, Sun S, Wang H, Hang H: **Comparison of autotransporter and ice nucleation protein as carrier proteins for antibody display on the cell surface of *Escherichia coli***. *Prog Biochem Biophys* 2013, **40**:1209–1219.

38. Verhoeven GS, Alexeeva S, Dogterom M, den Blaauwen T: **Differential bacterial surface display of peptides by the transmembrane domain of OmpA**. *PLoS One* 2009, **4**.

39. Samuelson P, Hansson M, Ahlborg N, Andreoni C, Gotz F, Bachi T, Thien Ngoc Nguyen, Binz H, Uhlen M, Stahl S: **Cell surface display of recombinant proteins on *Staphylococcus carnosus***. *J Bacteriol* 1995, **177**:1470–1476.

40. Binder U, Matschiner G, Theobald I, Skerra A: **High-throughput Sorting of an Anticalin Library via EspP-mediated Functional Display on the *Escherichia coli* Cell Surface**. *J Mol Biol* 2010, **400**:783–802.

41. Van Gerven N, Sleutel M, Deboeck F, De Greve H, Hernalsteens JP: **Surface display of the receptor-binding domain of the F17a-G fimbrial adhesin through the autotransporter AIDA-I leads to permeability of bacterial cells**. *Microbiology* 2009, **155**:468–476.

42. Fleetwood F, Devoogdt N, Pellis M, Wernery U, Muyldermans S, Ståhl S, Löfblom J: **Surface display of a single-domain antibody library on Gram-positive bacteria.** *Cell Mol life Sci* 2013, **70**:1081–93.

43. Salema V, Marín E, Martínez-Arteaga R, Ruano-Gallego D, Fraile S, Margolles Y, Teira X, Gutierrez C, Bodelón G, Fernández LÁ: **Selection of single domain antibodies from immune libraries displayed on the surface of *E. coli* cells with two β-domains of opposite topologies.** *PLoS One* 2013, **8**:e75126.

44. Pardon E, Laeremans T, Triest S, Rasmussen SGF, Wohlkönig A, Ruf A, Muyldermans S, Hol WGJ, Kobilka BK, Steyaert J: **A general protocol for the generation of Nanobodies for structural biology.** *Nat Protoc* 2014, **9**:674–93.

45. Brurberg MB, Eijsink VGH, Nes IF: **Characterization of a chitinase gene (chiA) from *Serratia marcescens* BJL200 and one-step purification of the gene product**. *FEMS Microbiol Lett* 1994, **124**:399–404.

46. Horn SJ, Sørbotten A, Synstad B, Sikorski P, Sørlie M, Vårum KM, Eijsink VGH: **Endo/exo mechanism and processivity of family 18 chitinases produced by *Serratia marcescens***. *FEBS J* 2006, **273**:491–503.

47. Yang C, Zhao Q, Liu Z, Li Q, Qiao C, Mulchandani A, Chen W: **Cell surface display of functional macromolecule fusions on *Escherichia coli* for development of an autofluorescent whole-cell biocatalyst**. *Environ Sci Technol* 2008, **42**:6105–6110.

48. Shi H, Wen Su W: **Display of green fluorescent protein on *Escherichia coli* cell surface**. *Enzyme Microb Technol* 2001, **28**:25–34.

49. Kang'ethe W, Bernstein HD: **Charge-dependent secretion of an intrinsically disordered protein via the autotransporter pathway.** *Proc Natl Acad Sci U S A* 2013, **110**:E4246–55.

50. Heggeset TMB, Kucharova V, Naerdal I, Valla S, Sletta H, Ellingsen TE, Brautaset T: **Combinatorial mutagenesis and selection of improved signal sequences and their application for high-level production of translocated heterologous proteins in *Escherichia coli***. *Appl Environ Microbiol* 2013, **79**:559–568.

51. Jarmander J, Janoschek L, Lundh S, Larsson G, Gustavsson M: **Process optimization for increased yield of surface-expressed protein in *Escherichia coli***. *Bioprocess Biosyst Eng* 2014, **37**:1685–1693.

52. Bendtsen JD, Nielsen H, Widdick D, Palmer T, Brunak S: **Prediction of twin-arginine signal peptides.** *BMC Bioinformatics* 2005, **6**:167.

53. Cavaleiro AM, Kim SH, Seppälä S, Nielsen MT, Nørholm MHH: **Accurate DNA Assembly and Genome Engineering with Optimized Uracil Excision Cloning**. *ACS Synth Biol* 2015, **4**:1042–1046.

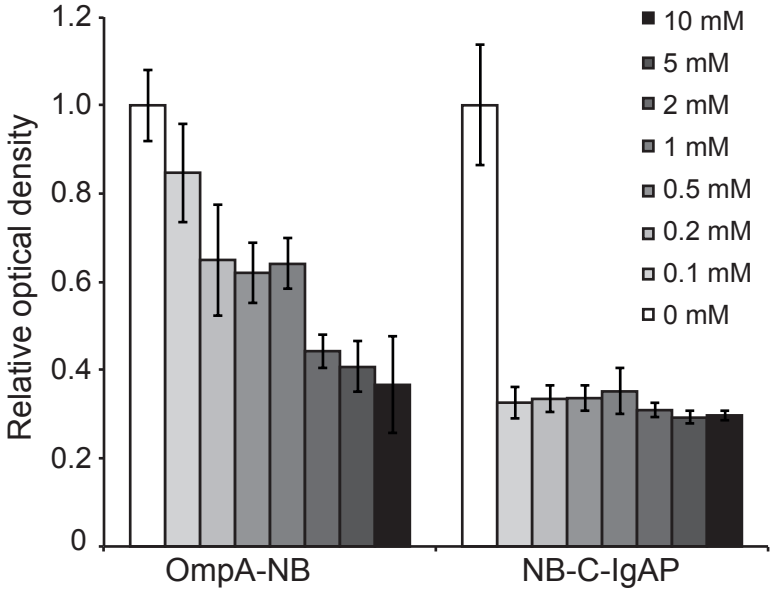
54. Drew DE, von Heijne G, Nordlund P, de Gier JW: **Green fluorescent protein as an indicator to monitor membrane protein overexpression in *Escherichia coli*.** *FEBS Lett* 2001, **507**:220–4.

55. Waldo GS, Standish BM, Berendzen J, Terwilliger TC: **Rapid protein-folding assay using green fluorescent protein.** *Nat Biotechnol* 1999, **17**:691–5.

56. Kim SH, Cavaleiro AM, Rennig M, Nørholm MHH: **SEVA linkers: a versatile and automatable DNA backbone exchange standard for synthetic biology**. *ACS Synth Biol* 2016:acssynbio.5b00257.

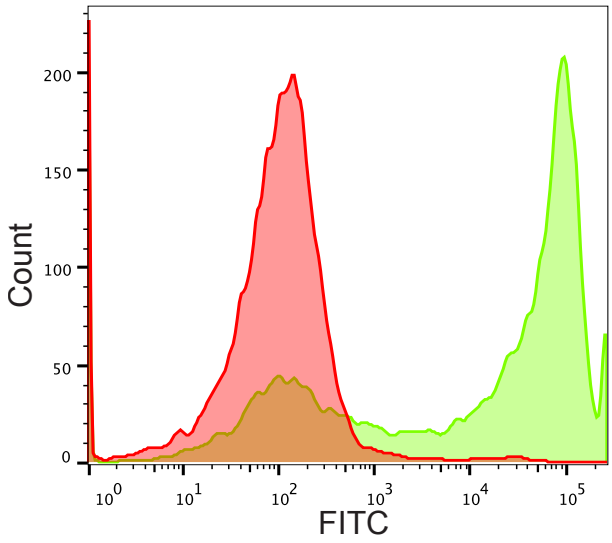
Additional files

Additional file 1: Figure S1



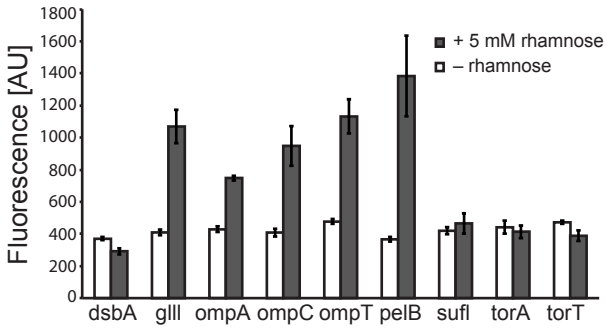
Cells expressing the NB-C-IgAP construct are negatively affected by induction of protein production. When induced with rhamnose of varying concentration, optical density of NB-C-IgAP cultures is decreasing drastically, and to a similar level independent of inducer concentration. LppOmpA- NB-producing cultures are also affected by induction, but less dramatically and in a stepwise manner. Values are normalised to the average of uninduced cells, biological triplicates, standard error.

Additional file 2: Figure S2



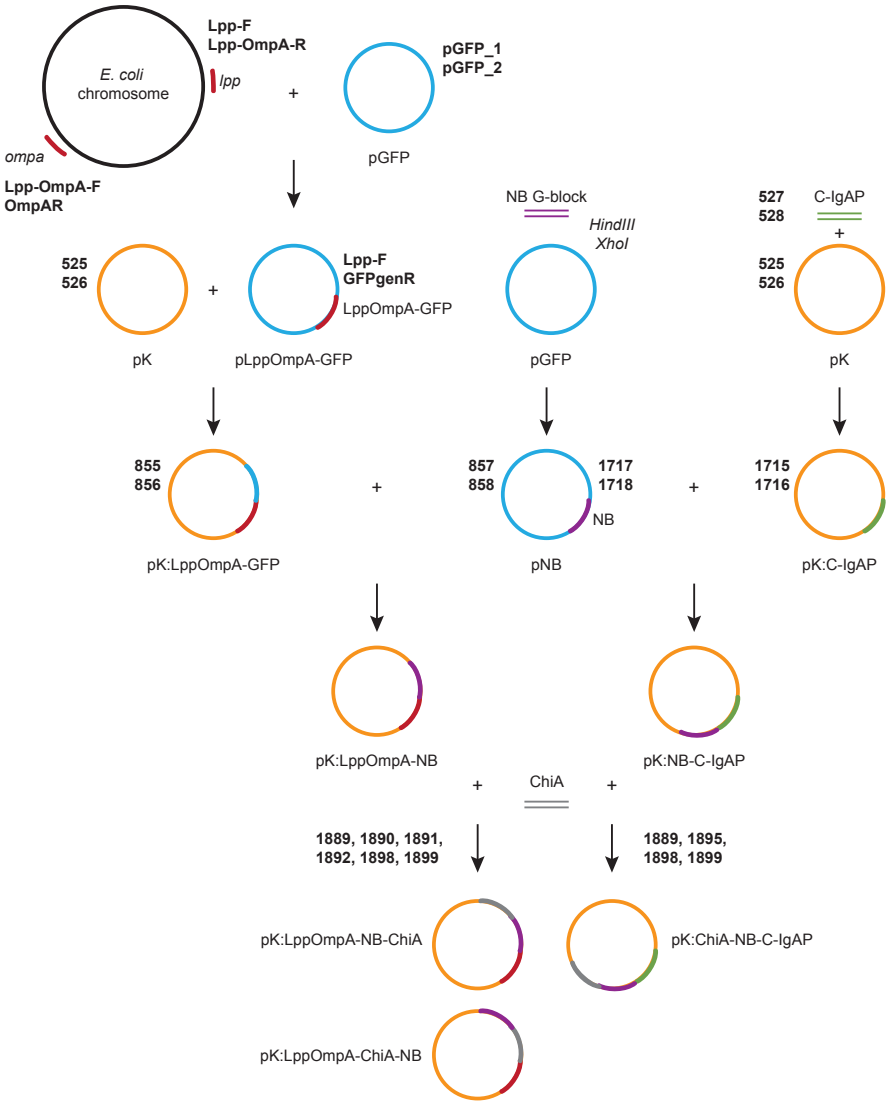
Population distribution with the P_{trc} promoter. The LppOmpA-NB fusion was cloned into a vector containing the IPTG-inducible P_{trc} promoter and subsequently assayed according to the described NB:GFP procedure, followed by flow cytometry analysis. As depicted in the figure, a large majority of cells were fluorescent.

Additional file 3: Figure S3



Variation in display signal when fusing OmpA-ChiA-NB to a set of 9 different signal peptides. The OmpA-ChiA-NB fusion protein was directed to the cell surface by different signal peptides, leading to varying levels of surface display. Values are averages of three biological replicates, error bars standard errors.

Additional file 4: Figure S4



Overview of plasmid construction. Oligos are given in bold font, plasmid names in regular font under the plasmid. Colours show from which source each fragment is amplified. The pK:LppOmpA-NB plasmid was made in several steps, starting with amplification of LppOmpA from the *E. coli*

chromosome, whereas pK:C-IgAP was created in one step. The nanobody sequence was ordered as a G- block and restriction cloned into pGFP. Plasmids with Chitinase A were made by cloning of the ChiA gene into pK:LppOmpA-NB and pK:NB-C-IgAP. Details are found in Gene and Vector design in the Methods section.

Additional file 5: Table S1.

Table S1. Oligos used in this study

Oligo ID	Sequence (5' to 3')
Lpp-F	ATATACCAUGAAAGCTACTAAACTGGTACTGG
Lpp-OmpA-R	aattccCUGATCGATTTTAGCGTTGCTGGAG
OmpAR	ACCCGGACCUCCGTTGTCCGGACGAGTGCCGATG
Lpp-OmpA-F	AGggaatUAACCCGTATGTTGGCTTTGAAATGGGTTAC
pGFP_1	AGGTCCGGGUATGAGCAAAGGAGAAGAACTTTTCAC
pGFP_2	ATGGTATAUCTCCTTCTTAAAGTTAAACAAAATTATTTCTAG
GFPgenR	ATCCTGGCUATTTGTAGAGCTCATCCATGCCATG
525	ATGGTATAUTCCTCCTGAATTTCATTACGAC
526	AGCCAGGAUAGAGTCGACCTGCAGGCATG
527	ATATACCAUGAAATACCTATTGCCTACGGCAG
528	ATCCTGGCUAGAAACGAATCTGTATTTTAATTTGTCCGGA TTTTTG
855	ATaCCCGGACCUCCGTTGTCCGGACGAGTG
856	AGCCAGGAUAGAGTCGACCTGCAGGCATG
857	AGGTCCGGGtAUGGCTCAGGTCCAACTGGTCG
858	ATCCTGGCUAGTGGTGGTGGTGGTGATGATG
1715	AGCAAGCTUGCGGCCCCACGTCCACCAACACCG
1716	AGCCATATUTGCCATACTAATTGCGGCT
1717	AATATGGCUCAGGTCCAACTGGTCGAATC
1718	AAGCTTGCUGCTAACCGTGACCTGCGT
1889	AGTCCAGGCAGUATGGCTCAGGTCCAACTGGTCGAATC
1890	accaggaccgcuTCCGTTGTCCGGACGAG
1891	AGTCCAGGCAGUtgaAGCCAGGATAGAGTCGACCTG
1892	accaggaccgcuGGCCGCAAGCTTGCTG
1895	accaggaccgcuAATTGCGGCTGAATTGTCGATCGG
1898	agcggtcctgguATGGCCGCGCCGGGCAAG
1899	actgcctggacuTTGAACGCCGGCCAGGCTGG
2148	ATATACCuATACCATGAAAGCTACTAAACTGG
2152	ACCTCAGCutcaGGCCGCAAGCTTGCT
2155	AGGTATAuCTCCTTCTTAAAGTTAAAC
2197	AGCTGAGGuCGCCTCAGC
2235	TAGGTACGAACTCGATTGACGgctcttctaccTCAGGCCGCAA GCTTGCTG
2236	TAGGTACGAACTCGATTGACGgctcttctaccCTATCCTGGCTA GAAACGAATCTG
2334	TACACGTACTTAGTCGCTGAAgctcttctatgCATCACCATCAC CATCACGCG
2539	TACACGTACTTAGTCGCTGAAgctcttctatgaTTAACCCGTATG TTGGCTTTGAA
2647	ATTCCGCuGGTGGTGCCGTTCTATAGCCATAGCATGCATC ACCATCACCATCACGC
2648	AGCGGAAuCGCGAACAGCAGTTTTTTCATGGTATATTCCT CCTGAATTTC

Additional file 6: Table S2.

Table S2. Plasmids used in this study

Plasmid	Reference
pK	[33]
pK:C-IgAP	This study
pK:LppOmpA-GFP	This study
pGFP	[54]
pKa:LppOmpA-NB	This study
pKa:NB-C-IgAP	This study
pKa:Lpp-OmpA-NB-chiA	This study
pKa:LppOmpA-chiA-NB	This study
pKa:chiA-NB-C-IgAP	This study
pCDF_sl3m:Ptrc	[56]
pCDF_sl3m:Ptrc-LppOmpA-NB	This study
pD881	DNA2.0
pD881:dsbAss-ChiA-NB-AT	This study
pD881:gIIIss-ChiA-NB-AT	This study
pD881:ompAss-ChiA-NB-AT	This study
pD881:ompCss-ChiA-NB-AT	This study
pD881:ompTss-ChiA-NB-AT	This study
pD881:pelBss-ChiA-NB-AT	This study
pD881:sufIss-ChiA-NB-AT	This study
pD881:torAss-ChiA-NB-AT	This study
pD881:torTss-ChiA-NB-AT	This study
pD881:dsbAss-OmpA-ChiA-NB	This study
pD881:gIIIss-OmpA-ChiA-NB	This study
pD881:ompAss-OmpA-ChiA-NB	This study
pD881:ompCss-OmpA-ChiA-NB	This study
pD881:ompTss-OmpA-ChiA-NB	This study
pD881:pelBss-OmpA-ChiA-NB	This study
pD881:sufIss-OmpA-ChiA-NB	This study
pD881:torAss-OmpA-ChiA-NB	This study
pD881:TorT-OmpA-ChiA-NB	This study
pKa_gIII_chi-NB-AT	This study